MTSummit 2025

MT SUMMIT Geneva 2025

Machine Translation Summit XX

Volume 1

Edited by:
Pierrette Bouillon, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, Sara Szoc

UNIVERSITÉ DE GENÈVE
FACULTÉ DE TRADUCTION ET D'INTERPRÉTATION

EUROPEAN ASSOCIATION FOR MACHINE TRANSLATION

June 23-27, 2025

Geneva, Switzerland

ISBN 978-2-9701897-0-1

Foreword from the General Chair

As president of the International Association for Machine Translation (IAMT) and General Chair of the 20th Machine Translation Summit, it is my utmost pleasure to write these opening words. Be most welcome to our MT Summit 2025!

The European Association for Machine Translation (EAMT) Executive Committee (EC) has been very busy. Mikel Forcada (treasurer) and Sara Szoc (secretary) have been tirelessly supporting all initiatives. Carolina Scarton and Sara Szoc took great care of our bursaries. Patrick Cadwell, André Martins, and Manuel Lardelli were our chairs for the Research Projects. Manuel Lardelli was also our policies chair, revisiting all our policies and contributing to inclusivity strategies. Our very own Mary Nurminen, chair of the bid proposals for our next events, has been busy selecting our next venue! EAMT 2026 venue will be disclosed in our closing ceremony in Geneva!

One of our core initiatives, the best thesis award – Rachel Badwen and Barry Haddow, chairs of the Best Thesis Award, had a very difficult time selecting a candidate, since the submissions were of very high quality. Our congratulations to Ricardo Rei's thesis "Robust, Interpretable and Efficient MT Evaluation with Fine-tuned Metrics" (Unbabel, INESC-ID, Instituto Superior Técnico, Portugal), supervised by Maria Luísa Torres Ribeiro Marques da Silva Coheur and Alon Lavie. We would also like to congratulate for the highly commended thesis of Sara Papi (University of Trento & Fondazione Bruno Kessler), entitled "Direct Speech Translation in Constrained Contexts: the Simultaneous and Subtitling Scenarios" and supervised by Marco Turchi and Matteo Negri.

EAMT, as full sponsor of the MT Marathon, would also like to thank the Institute of Formal and Applied Linguistics (ÚFAL), Charles University for organizing the 17th MT Marathon. The event included MT lectures and labs, covering the basics and tutorials; keynote talks from experienced researchers and practitioners; presentations of research and open source tools related to MT; and hacking projects to advance tools or research in one week or start new collaborations. A special thank you to Jindřich Helcl his commitment and passion for this event!

MT Summit 2025 will be a moment to celebrate our IAMT Award of Honour![1] We celebrate Professor Mikel Forcada, unanimously supported by all sister organizations (EAMT, AAMT, and AMTA), in recognition of his long-standing distinguished contribution to the EAMT and IAMT communities and for his impactful research on Machine Translation. Thank you for being an inspiration to us all!

Geneva, Switzerland, MT Summit 2025! Our conference will have a three-day, four-track programme put together by our chairs: Catarina Farinha and Marco Gaido (research: technical track chairs); Dorothy Kenny and Joke Adaems (research: translators & users track chairs); Samuel Läubli and Martin Volk (implementations & case studies track chairs); Miguel Esplà and Vincent Vandeghinste (products & projects track chairs) and François Yvon and Sheila Castilho (workshop and tutorial chairs). Our filters of quality and alignment! We really appreciate your work. We will continue with our tradition and also have a two-day workshops and tutorials event.

Our gratitude to all our keynotes speakers. Sarah Ebling, Full Professor of Language, Technology and Accessibility at the University of Zurich. Joss Moorkens, Associate Professor at the School of Applied Language and Intercultural Studies in Dublin City University (DCU). Eva Vanmassenhove, Assistant professor in the Department of Cognitive Science and Artificial Intelligence at Tilburg University (TiU). Our outstanding keynote speakers will demonstrate their extensive and global impactful work in translation studies and translation technologies, in a multidisciplinary motto which is the core of our community.

[1] https://eamt.org/iamt-award-of-honour/

MT Summit 2025 is the result of a very aligned, sharp, engaged, and hard working local organising team! What a diligent team! Our local co-chairs, Pierrette Bouillon, Johanna Gerlach, Sabrina Girletti and Lise Volkart (all from the University of Geneva, Switzerland) have put a lot of work in giving us a Geneva unforgettable event. To Sevita Caseres, Bastien David, Céline De Graaf, Julie Humbert-Droz, Rebeka Mali, Lucía Morado, Jonathan Mutal, Lucía Ormaechea, Aurélie Picton, Donatella Pulitano, Silvia Rodríguez, Raphael Rubino, Valentin Scourneau, Marianne Starlander, Irene Strasly, Nikolaos Tsourakis, Florine Voisard (all from the University of Geneva, Switzerland) and Rico Sennrich (University of Zurich, Switzerland), our deepest appreciation.

EAMT has been supported by generous sponsors in its initiatives along the years.[2] This year is no exception in a summit year! In fact, it is a very exceptional year in terms of sponsoring activities. Our gratitude to our Platinum sponsors who will also be giving a research oral presentation, BIG Language Solution, STAR, WIPO. Our Gold sponsor Systran by ChapsVision. Our Silver sponsors: Translated, Reverso, and Unbabel. To our Bronze sponsors: AppTek, CrossLang, TransPerfect, and Zoo Digital. To all our Supporter sponsors: Apertium, iguanodon.ai, prompsit, Springer Nature (our Supporter sponsor for the Best Paper award) and Supertext. Finally, to our Media sponsors, MultiLingual and Slator. Your support is vital in our efforts to give back to our community through grants and other initiatives.

A note still to all our IAMT members and our participants! Without you no effort would make sense! Let us take this opportunity to create scientific collaboration and give constructive feedback. To fully enjoy the conference, please check our Code of Conduct.[3] I'm looking forward to seeing you all and celebrating our community gathering!

Our sister organizations have been renewed with new board of Directors. The best wishes to AMTA's new board, represented by the President, Jay Marciano, and to the AAMT's Directors, Hisahiro Adachi, SunFlare Co., Ltd. (President of AAMT) and Masao Utiyama, National Institute of Information and Communications Technology, Japan (Vice President of AAMT). MT Summit 2027 will be held by AMTA! More soon!

It is our organisation's greatest wish to continue giving back to our community and to drive and be driven by our community's energy and enthusiasm. Reach out to us if you have new ideas or suggestions you would like to implement. We will try hard to accomplish it with you. Learn more about us.

Helena Moniz

President of the IAMT
General Chair of MT Summit 2025
University of Lisbon, Portugal

[2] https://mtsummit2025.unige.ch/sponsors.html
[3] https://mtsummit2025.unige.ch/about.html#codeOfConduct

Message from the Local Organising Committee

It is our great pleasure to welcome you to the Faculty of Translation and Interpreting (FTI) for this 20th edition of the MT Summit. We are particularly proud that for the first time in its history, the Summit is being hosted by a translation faculty, highlighting the importance of the human factor in today's technology. This is also a sign that technology has become an imperative in professional translation. Our faculty has long embraced this evolution, as illustrated by its translation technology department, first established back in the 1970s (first under the name of ISSCO, and then TIM). It was long spearheaded by Prof Maghi King, who, as some of you may recall, received the prestigious IAMT Award of Honour in 2005.

Our department has always been committed to building bridges between research in MT and professional translators. The conference taking place here today is further proof that this bridge is now well established and solid! The structure of the conference itself also reflects this dual focus, with two dedicated research tracks, one Technical, and the other for Translators and Users.

This year also brings an important new initiative: authors of papers involving computational experiments are encouraged to include sustainability reports. Most authors engaged with the initiative, reflecting the willingness of our community to embrace more transparent and thoughtful research practices.

We hope you will enjoy the rich and carefully curated program put together by our dedicated track chairs and made possible by the thorough work of our reviewers. We are also deeply grateful to our keynote speakers, as well as the organizers of the workshops and tutorials, whose contributions are crucial to the success of this conference.

We also want to thank our sponsors, more generous than ever before! Their presence is a strong indicator of the fruitful and trustworthy collaboration that exists between academia and industry in our field.

When we signed up to organise this conference, we had no idea of the summit that we would have to climb, nor how much determination, patience and endurance it would require of us. But thanks to our experience of the mountains, a dedicated team, and the valuable support of EAMT Executive Committee and previous organisers, we reached the (MT) Summit (almost) without problems. As in every climb, it is the strength of the team that gets you to the top!

We wish you an excellent MT Summit!

On behalf of the MT Summit 2025 Organising Committee:
Pierrette Bouillon
Johanna Gerlach
Sabrina Girletti
Lise Volkart

Department of Translation Technology (TIM)
Faculty of Translation and Interpreting
University of Geneva, Switzerland

Preface by the Programme Chairs

The **Research Technical track** received 57 submissions, out of which 28 were accepted, for an acceptance rate of 49%. 14 papers will be presented orally and the other 14 will be part of two poster sessions. The topics covered by the submitted papers include named entity aware translation, context-aware machine translation, domain-specific translation, multilingual and low-resource translation, and translation evaluation. We express our most heartfelt thanks to the 83 reviewers, who made this track possible, with a particular gratitude for the emergency reviewers who promptly accomplished their duties, enabling us to respect the timeline for author notification.

Catarina Farinha (Unbabel)
Marco Gaido (Fondazione Bruno Kessler, Italy)

The **Translators and Users track** initially received 28 submissions, of which 21 could be considered for this track, the other 7 covered more technical aspects of machine translation and were therefore considered for the Technical track instead. Of these 21, 19 were accepted (an acceptance rate of 90%, showing the overall high quality of submission to the track). As track chairs, we noticed a few trends in these accepted papers, and we tried to group the submissions in sessions accordingly. The large language model trend, established in earlier EAMT conferences, clearly continues. Large language models are used for literary translation (post-editing) and emergency response text translation, and there is a clear interest in how these technologies are currently being used by students as well as perceived by professionals. From the text types that are being studied, it is obvious that 'literary translation' is most strongly represented in this track, with 5 submissions covering the topic. This is particularly striking, given that this MT Summit is also hosting a dedicated workshop on Creative-text Translation and Technology. The intersection of creativity, literature and automatic translation has clearly arrived as a field of inquiry. We thank all PC members for their time and dedication in delivering insightful feedback, ensuring the quality of the submissions to this track. Special thanks to the emergency reviewers who helped us avoid any delays. You all made this conference possible.

Joke Daems (Ghent University, Belgium)
Dorothy Kenny (Dublin City University, Ireland)

The **Implementation and Case Studies track** received 12 submissions out of which 9 were accepted for presentation at the MT summit (6 talks and 3 posters). The papers cover a broad range of topics, e.g. speech translation, LLM-based translation, low-resource settings, productivity evaluation and translator satisfaction. We would like to express our gratitude and appreciation to our reviewers from academia and industry for their time and effort in commenting and grading the submissions.

Samuel Läubli (Textshuttle/Supertext, Switzerland)
Martin Volk (University of Zurich, Switzerland)

The **Products and Projects track** received 22 submissions, of which 20 have been accepted for a short, two-page description and a poster presentation at the conference. Our selection highlights a diverse range of products and projects created by our community, covering research projects and cutting-edge services and innovations from distinguished industry and research leaders. Expect a lively session filled with poster boasters and engaging poster presentations. We wish to thank the 26 members of the program committee for this track for their timely and thorough reviews.

Miquel Esplà-Gomis (University of Alicante, Spain)
Vincent Vandeghinste (KU Leuven, Belgium)

The **Workshop and Tutorials** received seven workshop proposals, five of which were finally selected: four are reiterations of workshops that have already been collocated with MT conferences in the past: these are the "2nd Workshop on Creative-text Translation and Technology" (CTT 2025), the 3rd "International Workshop on Gender-Inclusive Translation Technologies" (GITT 2025), the 3rd "International Workshop on Automatic Translation for Signed and Spoken Languages" (AT4SSL), and the 11th "Workshop on Patent and Scientific Literature Translation" (PSLT 2025). We are also happy to see the start of a hopefully equally successful new series, with the 1st "Workshop on Artificial Intelligence and Easy and Plain Language in Institutional Contexts" (AI & EL/PL). With the exception of PSLT, they will all run for a full day, on the 23rd or on the 24th of June. Five half-day tutorials were also submitted, and three will be offered to the participants: "Understanding Large Language Model-Generated Translations", "Leveraging Examples in Machine Translation", and "Best practices for data quality in human annotation of translation datasets". Our hope is that the choice between such diverse and exciting proposals will be a difficult one, and that these two pre-conference days will be as enjoyable and rewarding as possible, sparking new ideas, collaborations, and conversations in Geneva and beyond.

Sheila Castilho (Dublin City University, Ireland)
François Yvon (Sorbonne University, France)

EAMT 2024 Best Thesis Award (Anthony C. Clarke Award)

Six PhD theses defended in 2024 were received as candidates for the 2024 year edition of the EAMT Best Thesis Award, all of which were eligible. Eight external reviewers were recruited to examine and score the theses alongside five EAMT executive committee members. Each thesis was evaluated according to predefined criteria: how challenging the topic was, how relevant the results were to the MT field and the strength of its impact in terms of scientific publications. As in previous years, 2024 was another strong year for PhD theses in machine translation.

All PhD theses were of good quality, focused on interesting topics and were all highly appreciated by reviewers. A panel of two EAMT Executive Committee members (Barry Haddow and Rachel Bawden) was assembled to process the reviews and select a winner that was later ratified by the EAMT executive committee.

We are pleased to announce that the **winner of the 2024 edition of the EAMT Best Thesis Award is Ricardo Rei's thesis "Robust, Interpretable and Efficient MT Evaluation with Fine-tuned Metrics"** (Unbabel, INESC-ID, Instituto Superior Técnico, Portugal), supervised by Maria Luísa Torres Ribeiro Marques da Silva Coheur and Alon Lavie.

In addition, the committee judged that the thesis of **Sara Papi** (University of Trento & Fondazione Bruno Kessler) entitled "Direct Speech Translation in Constrained Contexts: the Simultaneous and Subtitling Scenarios" and supervised by Marco Turchi and Matteo Negri was **"highly commended"**.

The awardee will receive a prize of €500, together with an inscribed certificate. In addition, Dr. Rei will present a summary of their thesis at the 20th Machine Translation Summit in Geneva, Switzerland, receive complimentary membership to the EAMT in 2026 and will receive a travel bursary of €200.

Chairs of the EAMT Best Thesis Award 2024
Rachel Bawden, Inria, Paris, France
Barry Haddow, University of Edinburgh, UK

Organising Committee

General Chair

Helena Moniz, Universidade de Lisboa / INESC-ID, Portugal

Local Organising Committee

Pierrette Bouillon, University of Geneva, Switzerland
Johanna Gerlach, University of Geneva, Switzerland
Sabrina Girletti, University of Geneva, Switzerland
Lise Volkart, University of Geneva, Switzerland

Local Support Team

Sevita Caseres, University of Geneva, Switzerland
Bastien David, University of Geneva, Switzerland
Céline De Graaf, University of Geneva, Switzerland
Julie Humbert-Droz, University of Geneva, Switzerland
Rebeka Mali, University of Geneva, Switzerland
Lucía Morado Vázquez, University of Geneva, Switzerland
Jonathan Mutal, University of Geneva, Switzerland
Lucía Ormaechea Grijalba, University of Geneva, Switzerland
Aurélie Picton, University of Geneva, Switzerland
Donatella Pulitano, University of Geneva, Switzerland
Silvia Rodríguez Vázquez, University of Geneva, Switzerland
Valentin Scourneau, Université de Mons
Marianne Starlander, University of Geneva, Switzerland
Irene Strasly, University of Geneva, Switzerland
Nikolaos Tsourakis, University of Geneva, Switzerland
Florine Voisard, University of Geneva, Switzerland

Publications Chair

Raphael Rubino, University of Geneva, Switzerland
Rico Sennrich, University of Zurich, Switzerland

Track Chair: Research Technical

Catarina Farinha, Unbabel, Portugal
Marco Gaido, Fondazione Bruno Kessler, Italy

Track Chair: Research Translators and Users

Joke Daems, Ghent University, Belgium
Dorothy Kenny, Dublin City University, Ireland

Track Chair: Implementations and Case Studies

Samuel Läubli, Textshuttle/Supertext, Switzerland
Martin Volk, University of Zurich, Switzerland

Track Chair: Products and Projects

Miquel Esplà-Gomis, University of Alicante, Spain
Vincent Vandeghinste, KU Leuven, Belgium

Workshops and Tutorials Chair

Sheila Castilho, Dublin City University, Ireland
François Yvon, Sorbonne University, France

Programme Committee

Track: Research Technical

Benyamin Ahmadnia	UC Davis
Dr Khetam Al Sharou	Imperial College London
Àlex R. Atrio	HEIG-VD / HES-SO & EPFL
Vicent Briva-Iglesias	SFI CRT D-REAL, Dublin City University
José G. C. de Souza	Unbabel
Vera Cabarrão	Unbabel Lda.; INESC-ID
Michael Carl	Kent State University
Luisa Coheur	INESC-ID
Mattia Antonino Di Gangi	AppTek
Siddharth Divi	SSN College of Engineering
Konstantin Dranch	Custom.MT
Kevin Duh	Johns Hopkins University
Hiroshi Echizenya	Hokkai-Gakuen University
Carlos Escolano	UPC - BSC
Miquel Esplà-Gomis	Universitat d'Alacant
Mikel Forcada	Universitat d'Alacant
Javier García Gilabert	Barcelona Super Computing Center (BSC)
Cyril Goutte	National Research Council Canada
Barry Haddow	University of Edinburgh
Rejwanul Haque	South East Technological University
Iikka Hauhio	University of Helsinki, Kielikone Oy
Javier Iranzo-Sánchez	Universitat Politecnica de Valencia
Josef Jon	Charles University
Swarang Joshi	IIIT Hyderabad
Alina Karakanta	Leiden University
Maria Kunilovskaya	University of Saarland
Natalie Kübler	University of Paris
Gorka Labaka	University of the Basque Country
Tsz Kin Lam	University of Edinburgh
Ekaterina Lapshinova-Koltunski	University of Hildesheim
Yves Lepage	Waseda University
Qun Liu	Huawei Noah's Ark Lab
John Mendonca	INESC-iD
Miguel Menezes	Lisboa, Inesc-ID, Unbabel
Thomas Moerman	Ghent University
Kenton Murray	Johns Hopkins University
Jonathan Mutal	UNIGE
Masaaki Nagata	NTT
Artur Nowakowski	Laniqo / Adam Mickiewicz University
Constantin Orasan	University of Surrey
David Orrego-Carmona	University of Warwick
Antonio Pareja-Lora	ATLAS (UNED) / FITISPos (UAH) / DMEG (UdG, México) / DSIC, ILSA (UCM)
Seong-Bae Park	Kyung Hee University
Patrícia Pereira	Instituto Superior Técnico
Andrea Piergentili	University of Trento
Esther Ploeger	Aalborg University

David Ponce	Vicomtech
Andrei Popescu-Belis	HEIG-VD / HES-SO
Maja Popovic	ADAPT Centre @ DCU
Bo Ren	Microsoft
Fatiha Sadat	UQAM
Beatrice Savoldi	Fondazione Bruno Kessler
Yves Scherrer	University of Oslo
Dimitar Shterionov	Tilburg University
Michel Simard	National Research Council Canada (NRC)
Patrick Simianer	Lilt, Inc.
Sokratis Sofianopoulos	ILSP / Athena R.C.
Rubén Solera-Ureña	INESC-ID Lisboa
Rui Sousa-Silva	University of Porto
Felix Stahlberg	Google Research
Katsuhito Sudoh	Nara Women's University
Marek Suppa	Comenius University in Bratislava
Felipe Sánchez-Martínez	Universitat d'Alacant
Marina Sánchez-Torrón	Smartling
Aleš Tamchyna	Phrase a.s.
Antonio Toral	Universitat d'Alacant
Marco Turchi	Zoom
Jannis Vamvas	University of Zurich
Vincent Vandeghinste	Instituut voor de Nederlandse Taal, Leiden // Centre for Computational Linguistics, KU Leuven
David Vilar	Google
Taro Watanabe	Nara Institute of Science and Technology
Guillaume Wisniewski	LLF - Université de Paris
Tong Xiao	Northeastern University (CN)
Jinan Xu	Beijing Jiaotong University
Rik van Noord	University of Groningen

Track: Research Translators and Users

Sergi Alvarez-Vidal	UPF
Fabio Alves	UFMG
Nora Aranberri	University of the Basque Country
Lynne Bowker	Université Laval
Vicent Briva-Iglesias	SFI CRT D-REAL, Dublin City University
Patrick Cadwell	Dublin City University
Dragos Ciobanu	University of Vienna
Helle Dam Jensen	Aarhus University
Christophe Declercq	Utrecht University
Silvana Deilen	University of Hildesheim
Félix Do Carmo	CTS - University of Surrey
Aletta G. Dorst	Leiden University
Maria Fernandez-Parra	Swansea University
Federico Gaspari	ADAPT Centre, Dublin City University
Ana Guerberof Arenas	University of Groningen
Sari Hokkanen	Tampere University
Maarit Koponen	University of Eastern Finland
Ekaterina Lapshinova-Koltunski	University of Hildesheim
Manuel Lardelli	University of Graz

Rudy Loock	Université de Lille, France, & CNRS "Savoirs, Textes, Langage" research unit
Lieve Macken	Ghent University
Joss Moorkens	Dublin City University
Lucas N Vieira	University of Bristol
Masaaki Nagata	NTT
Mary Nurminen	University of Eastern Finland and Tampere University
Antoni Oliver	Universitat Oberta de Catalunya
Constantin Orasan	University of Surrey
David Orrego-Carmona	University of Warwick
John Ortega	Columbia and New York Universities
Jun Pan	Hong Kong Baptist University
Celia Rico	Universidad Complutense de Madrid
Akiko Sakamoto	Kansai University
Vilelmini Sosoni	Ionian University
Sanjun Sun	Beijing Foreign Studies University
María Del Mar Sánchez Ramos	Universidad de Alcala
Susana Valdez	Leiden University Centre for Linguistics
Kirti Vashee	Translated Srl
Mihaela Vela	Universität des Saarlandes
Callum Walker	University of Leeds

Track: Implementations and Case Studies

Chantal Amrhein	Supertext
Thomas Brovelli	Google
Oliver Czulo	Universität Leipzig
Marcello Federico	AWS AI Labs
Mark Fishel	University of Tartu
Tim Graf	Supertext
Ana Guerberof Arenas	University of Groningen
Silvia Hansen-Schirra	Johannes Gutenberg-Universität Mainz
Martin Kappus	Zürcher Hochschule für Angewandte Wissenschaften
Judith Klein	STAR Group
Maarit Koponen	University of Eastern Finland
Alon Lavie	Phrase
Christian Lieske	SAP
Helena Moniz	University of Lisbon
Mary Nurminen	University of Eastern Finland and Tampere University
Carla Parra Escartín	RWS Language Weaver
Matiss Rikters	Tilde
Florian Schottmann	Supertext
Sara Szoc	CrossLang
Carlos Teixeira	Universitat Rovira i Virgili
Jannis Vamvas	University of Zurich
Masaru Yamada	Rikkyo University
Maike Züfle	Karlsruher Institut für Technologie

Track: Products and Projects

Sergi Alvarez-Vidal	UPF
Eleftherios Avramidis	German Research Center for Artificial Intelligence (DFKI)
Romane Bodart	Université catholique de Louvain

Pedro Luis Díez-Orzas	Linguaserve I.S. S.A.
Judith Klein	STAR Group
Rebecca Knowles	National Research Council Canada
Ekaterina Lapshinova-Koltunski	University of Hildesheim
Manuel Lardelli	University of Graz
Marie-Aude Lefer	Université catholique de Louvain
Lieve Macken	Ghent University
Maite Melero	UPF
Yasmin Moslem	ADAPT Centre, Dublin City University
Vlad Niculae	Instituto de Telecomunicacoes, Lisboa
Mary Nurminen	University of Eastern Finland and Tampere University
Antoni Oliver	Universitat Oberta de Catalunya
Juan Antonio Pérez-Ortiz	Universitat d'Alacant, Departament de Llenguatges i Sistemes Informàtics
Shenbin Qian	University of Surrey
Felipe Sánchez-Martínez	Universitat d'Alacant
Arda Tezcan	Ghent University
Antonio Toral	Universitat d'Alacant
Daniel Torregrosa	WIPO
Tom Vanallemeersch	CrossLang NV
Bram Vanroy	Instituut voor de Nederlandse Taal
Rik van Noord	University of Groningen

Keynote Talk

Sign Language Machine Translation

Sarah Ebling
University of Zurich (UZH)

Abstract: In this talk, I will highlight the challenges of automatic translation between spoken languages and sign languages, touching on the topics of representation, data, and ethics. Additionally, I will introduce preprocessing tasks and discuss their state of the art. I will present research conducted in our group in the different areas.

Bio: Sarah Ebling is Full Professor of Language, Technology and Accessibility at the University of Zurich. Based in the field of computational linguistics, her research focuses on language-based assistive technologies in the context of persons with disabilities. Specifically, Sarah Ebling's research takes place in the context of deafness and hearing impairment, blindness and visual impairment, cognitive impairment, and language disorders. She is conducting research on sign language technologies, automatic text simplification, technologies for the audio description process, and computer-aided language sample analysis. Sarah Ebling is involved in international and national projects and is the PI of a large-scale Swiss innovation project entitled Inclusive Information and Communication Technologies"(2022-2026; https://www.iict.uzh.ch/).

Keynote Talk

Losing Our Tail – Again: Unnatural Selection and Translation Technologies

Eva Vanmassenhove
Tilburg University (TiU)

Abstract: Language is humanity's primary tool to preserve and transmit knowledge, evolving alongside and with cultural technologies. Today, multilingual large language models (LLMs) represent the latest leap. Emerging evidence, however, suggests that LLMs might subtly (or not so subtly) distort language over time, amplifying frequent patterns while eroding linguistic richness, a phenomenon linked to *model collapse* which had already been observed in Neural Machine Translation (NMT) systems even before it was formally named. Unlike the visible artefacts that have already been observed in the AI-generated images created by computer vision models, linguistic shifts, such as the loss of the long tails of language, risk going unnoticed. Yet, they mayhave profound implications for language, translation, diversity, and the integrity of communication across different languages. This keynote will explore these ideas and connect them to specific translation issues, asking: What is (or will be) at stake when our world of words becomes increasingly shaped by multilingual LLMs.

Bio: Eva Vanmassenhove is a researcher specializing in Machine Translation and Language Technology, with a strong focus on tackling gender and algorithmic biases in translation systems. She earned her PhD from Dublin City University and now serves as an assistant professor in the Department of Cognitive Science and Artificial Intelligence at Tilburg University (TiU). At TiU, she contributes to the Computation and Psycholinguistics Research unit and the Inclusive and Sustainable Machine Translation Research Line. Her work aims to enhance machine translation by addressing biases, especially in gender representation, while preserving linguistic richness.

Keynote Talk

Ethics and MT Evaluation: An Exploded View

Joss Moorkens
Dublin City University (DCU)

Abstract: This talk reflects on ethical issues with MT using LLMs, looking particularly at a recent evaluation study in the medical domain. This study, and the potential for its findings to be used as a basis for action, bring abstract ethical issues into focus. More broadly, the heightened attention and potential for impact of MT and LLM research brings an added sense of responsibility for researchers, although this might be balanced with opportunities to contribute to the common good.

Bio: Joss Moorkens is an Associate Professor at the School of Applied Language and Intercultural Studies in Dublin City University (DCU), Science Lead at the ADAPT Centre, and member of DCU's Institute of Ethics and Centre for Translation and Textual Studies. He has published over 60 articles and papers on the topics of translation technology interaction and evaluation, translator precarity, and translation ethics. He is General Co-Editor of the journal Translation Spaces with Prof. Dorothy Kenny, co-editor of a number of books and journal special issues, and co-author of the textbooks Translation Tools and Technologies (Routledge 2023) and Automating Translation (Routledge 2024). He sits on the board of the European Masters in Translation Network.

Tutorial

Understanding Large Language Model-Generated Translations: How Can They Adapt to Different Translation Specifications and Pass the Translation Turing Test?

Longhui Zou[1], **Michael Carl**[2], **Alan Melby**[3], **Brandon Torruella**[4], **Masaru Yamada**[5]
[1]University of Montana, [2]Kent State University - CRITT, [3]International Federation of Translators,
[4]Brigham Young University, [5]Rikkyo University

Abstract: This tutorial explores the practical application of the Translation Turing Test (TTT) within today's evolving generative AI landscape, addressing the growing need for human-centered approaches to translation project management and machine translation evaluation. While substantial research has examined large language models (LLMs)' translation quality, little attention has been paid to their potential in managing the complex human interactions that characterize real-world translation project negotiations.

The TTT is a translation-specific adaptation of the classic Turing Test, evaluating whether a machine-managed translation project can successfully imitate a professional human project manager. In the TTT, a requester interacts with both human and computer systems to negotiate translation specifications and conduct a complete translation project. The machine passes if the requester cannot distinguish between the two managers more than 30% of the time.

This half-day tutorial guides participants through current language industry practices and the three major TTT components: specification negotiation, target text quality assessment, and complaint negotiation. By comparing three translation project cycles (managed by a human professional, a trained amateur, and a generative AI agent), we evaluate whether LLM-powered agents can handle complex coordination tasks characteristic of language service providers.

The program includes four sessions: introduction to the TTT, demonstration of requester-provider negotiations, translation quality evaluation including MQM customization and syntactic complexity analysis, and complaint negotiations. Participants gain both theoretical understanding and practical experience assessing the feasibility of integrating LLMs into real-world translation projects that support or enhance human project managers' roles.

Tutorial

Leveraging Examples in Machine Translation: A Guide to Retrieval and Integration Strategies

Maxime Bouthors[1], **Josep Maria Crego**[2]
[1]ISIR - Sorbonne Université - CNRS, [2]SYSTRAN by ChapsVision

Abstract: Retrieval-Augmented Generation (RAG) systems are growing popular in the era of Large Language Models (LLM). Nonetheless, retrieval augmentation has a long time story tied to Machine Translation (MT). This tutorial aims to put in perspective the various techniques used to (1) retrieve relevant examples for databases; (2) integrate them into MT models. We will uncover how the selection of examples can be performed (fuzzy matching, cross-lingual retrieval), some of the model architectures (edit-based models, augmented encoder-decoder generation models, LLMs), as well as how the augmentation affects the output. The target audience are academics and industry professionals wishing to incorporate examples to improve their translation quality.

Tutorial

Best Practices for Data Quality in Human Annotation of Translation Datasets

Marina Sánchez Torrón[1], **Jennifer Wong**[1]
[1]Smartling

Abstract: High-quality human annotations are essential for developing and evaluating machine learning (ML) models. However, annotation is a complex task, and creating reliable annotation datasets requires addressing multiple challenges. This tutorial provides comprehensive guidance on best practices for managing data quality in human annotation of translation datasets using the Multidimensional Quality Metrics (MQM) framework. Drawing from both academic research and industry experience, we cover the complete annotation lifecycle: from initial setup and annotator management to quality evaluation and improvement strategies. Through theoretical foundations and a practical demonstration, participants will learn concrete guidelines they can apply to create more reliable and consistent annotation datasets.

Table of Contents

Best Thesis Award

Robust, interpretable and efficient MT evaluation with fine-tuned metrics

Ricardo Rei
Instituto Superior Técnico, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
INESC-ID HLT lab, Rua Alves Redol, 9 1000-029 Lisboa, Portugal
Unbabel Research, R. Castilho 52, 1250-069 Lisboa, Portugal
Supervisors: Luísa Coheur and Alon Lavie
ricardo.rei@tecnico.ulisboa.pt

With the increasing need for Machine Translation (MT) in a world which is becoming globalized, there is also an increasing need to constantly evaluate the quality of the produced translations. This evaluation can be achieved through human annotators performing quality assessments or by employing automatic metrics. While human evaluation is preferred, it is expensive and time-consuming. Consequently, over the past decade, MT progress has primarily been measured using automatic metrics that assess lexical similarity against reference translations. However, numerous studies have demonstrated that lexical-based metrics do not correlate well with human judgments, casting doubt on the reliability of research in MT. Motivated by these challenges, the main goal of this thesis was to improve the current state of MT evaluation by developing new automatic metrics that satisfy four criteria: 1) strong correlation with human judgments, 2) robustness across different domains and language pairs, 3) interpretability, and 4) efficiency.

Based on recent advancements in cross-lingual language modeling, we hypothesize that a supervised metric incorporating the source-language input into the evaluation process will produce a more accurate MT evaluation. To validate this hypothesis, we introduce COMET (Crosslingual Optimized Metric for Evaluation of Translation), a neural framework for training multilingual MT evaluation models that serve as metrics. Models developed within the COMET framework are trained to predict human judgments of MT quality, such as *Direct Assessments* (DA), *Multidimensional Quality Metrics* (MQM), or *Human-mediated Translation Edit Rate* (HTER). Our results demonstrate that metrics developed within our framework achieve state-of-the-art correlations with human judgments across various domains and language pairs.

Nevertheless, lexical metrics still possess redeeming qualities in terms of interpretability and lightweight nature. In contrast, fine-tuned neural metrics like COMET are considered "slow black-boxes". To address this, we employ neural explainability methods to reveal that these metrics leverage token-level information directly associated with translation errors. We showcase their effectiveness for interpreting state-of-the-art fine-tuned neural metrics by comparing token-level neural saliency maps with MQM annotations. Additionally, we present several experiments aimed at reducing the computational cost and model size of COMET while maintaining its state-of-the-art correlation with human judgments, thus bridging the performance gap between lexical and model-based metrics. That work, titled COMETINHO: THE LITTLE METRIC THAT COULD, was recognized with the Best Paper Award at EAMT 2022.

Realizing that system-level MT metrics alone are insufficient for comprehensive evaluation, this thesis also presents MT-TELESCOPE, a contrastive analysis tool that provides fine-grained segment-level insights into MT quality. By identifying the factors behind system performance, MT-TELESCOPE enables a deeper understanding of translation accuracy at the phenomenon level (e.g., named entities).

Over the past years, my thesis work has significantly influenced the field, inspiring research on quality-aware decoding – a paradigm that closely aligns with recent advances in test-time compute for large language models. By introducing high-performing, interpretable, and efficient evaluation metrics, my thesis work represents a substantial step forward in MT evaluation and has set a new standard for assessing translation quality. Receiving the EAMT 2022 Best Paper Award along with the Best Thesis Award at EAMT 2024 is a great honor and further solidifies the strength and recognition of my work in MT by the EAMT organizers.

Proceedings of Machine Translation Summit XX Volume 1, pages 1–1, June 23-27, 2025

Thesis Title: "*Direct Speech Translation in Constrained Contexts: the Simultaneous and Subtitling Scenarios*"

Sara Papi
University of Trento and Fondazione Bruno Kessler
spapi@fbk.eu

ADVISORS:

- Marco Turchi (Zoom Video Communications)
- Matteo Negri (Fondazione Bruno Kessler)

EXTERNAL REVIEWERS:

- Claudio Fantinuoli (University of Mainz and KUDO Inc.)
- Juan Pino (Meta AI)

SUMMARY:

1 Motivation

The shift to online communications in various sectors like business, education, and entertainment has highlighted the need for effective language translation to enable seamless interaction among users with diverse linguistic and accessibility needs. Speech-to-text translation (ST) emerges as a core technology for overcoming language barriers and facilitating communication by converting spoken words into another language, offering a natural understanding of language. However, developing ST systems is challenging due to the inherent complexities of speech, such as variations in accents, speaking rates, and background noise. These challenges are further complicated by constraints such as time (e.g., output latency), space (e.g., characters to be displayed on the screen), computational resources (e.g., using CPUs or GPUs), or limited data availability (e.g., low resource languages).

2 Research Questions

The objective of ST is to achieve the highest quality of automatic textual translations. However, many applications require more than just high translation quality. When additional constraints are present, the challenge becomes balancing translation quality with these specific requirements. This PhD thesis focuses on two constrained scenarios: simultaneous speech translation and automatic subtitling. Both tasks are of significant scientific and industrial interest.

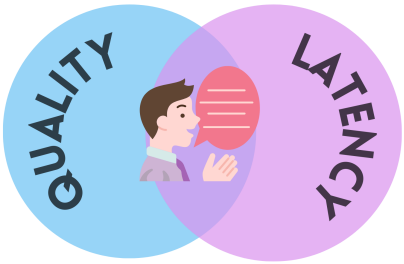


Figure 1: Simultaneous Translation Constraints.

2.1 Simultaneous Translation

Simultaneous Speech Translation (SimulST) aims to minimize latency–the delay from when an utterance is spoken in the source language to when it is translated into the target language. This requires translations to be displayed promptly and aligned with the natural pace of speech. Balancing translation quality and latency is essential for user comprehension and experience (Figure 1). Current SimulST systems face challenges in achieving this balance and often require complex training procedures with multiple training stages and sometimes the need to develop several models for different latency requirements. This PhD research explored whether existing ST systems possess intrinsic knowledge that can be leveraged for real-time applications without complex, ad-hoc training procedures. The main research questions were:

- Are complex training procedures necessary for SimulST?
- Can the knowledge acquired by standard ST models be used to guide them during simultaneous inference?

Proceedings of Machine Translation Summit XX Volume 1, pages 2–3, June 23-27, 2025

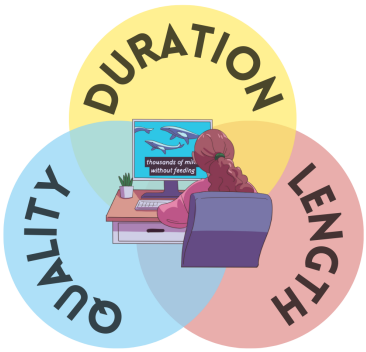


Figure 2: Automatic Subtitling Constraints.

2.2 Automatic Subtitling

Automatic Subtitling translates spoken dialogue in audiovisual media into text, which has to conform to spatial constraints (subtitle length) and temporal constraints (synchronization with audiovisual content). Long subtitles may overwhelm viewers, while short ones risk losing information; thus, proper subtitle length and synchronization ensure they remain on screen long enough to be read without disrupting the video's flow (Figure 2). In this scenario, prosody and speech cues are crucial for subtitle segmentation and timing, but current cascade architectures lose this information. Therefore, this PhD research aimed to leverage direct models that have direct access to this information, addressing two key questions:

- Is there a way to exploit prosody and speech cues to build automatic subtitling datasets starting from already existing ST corpora, overcoming data scarcity?
- Can a direct ST model produce fully segmented and timed subtitles?

3 Contibutions

3.1 Simultaneous Translation

In SimulST, the goal was to assess if standard ST systems could be repurposed for real-time use by leveraging their intrinsic knowledge, advocating a paradigm shift in model development. The contributions can be summarized in the findings below:

- Standard ST systems used for simultaneous inference achieve competitive or superior quality and latency compared to those ad-hoc trained for the tasks (Papi et al., 2022a);
- Intrinsic knowledge, particularly cross-attention, can be effectively used for SimulST, resulting in low-latency translation with minimal computational costs (Papi et al., 2023b);
- Using cross-attention for aligning speech and translation to guide simultaneous inference achieves an optimal balance between quality and latency (Papi et al., 2023c).

3.2 Automatic Subtitling

In Automatic Subtitling, the goal was to use direct systems, able to exploit speech cues, for subtitle segmentation and to generate complete subtitles. Specifically, the main findings are:

- To cope with data scarcity, direct multilingual multimodal models, which utilize both audio and textual cues to identify optimal segmentation points, revealed their effectiveness in automatic subtitle segmentation, delivering performance comparable to gold segmentation (Papi et al., 2022b);
- Direct ST models demonstrate the capability of generating full subtitles, which consist of segmented translations with corresponding timestamps, showing competitive performance against existing production tools (Papi et al., 2023a).

References

Sara Papi, Marco Gaido, Alina Karakanta, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2023a. Direct speech translation for automatic subtitling. *Transactions of the Association for Computational Linguistics*, 11:1355–1376.

Sara Papi, Marco Gaido, Matteo Negri, and Marco Turchi. 2022a. Does simultaneous speech translation need simultaneous models? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 141–153, Abu Dhabi, United Arab Emirates.

Sara Papi, Alina Karakanta, Matteo Negri, and Marco Turchi. 2022b. Dodging the data bottleneck: Automatic subtitling with automatically segmented ST corpora. 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 2: Short Papers)*, pages 480–487, Online only.

Sara Papi, Matteo Negri, and Marco Turchi. 2023b. Attention as a guide for simultaneous speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13340–13356, Toronto, Canada.

Sara Papi, Marco Turchi, and Matteo Negri. 2023c. Alignatt: Using attention-based audio-translation alignments as a guide for simultaneous speech translation. In *Interspeech 2023*, pages 3974–3978.

Research – Technical

Investigating Length Issues in Document-level Machine Translation

Ziqian Peng[1,2], Rachel Bawden[2], 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

Transformer architectures are increasingly effective at processing and generating very long chunks of texts, opening new perspectives for document-level machine translation (MT). In this work, we challenge the ability of MT systems to handle texts comprising up to several thousands of tokens. We design and implement a new approach designed to precisely measure the effect of length increments on MT outputs. Our experiments with two representative architectures unambiguously show that (a) translation performance decreases with the length of the input text; (b) the position of sentences within the document matters, and translation quality is higher for sentences occurring earlier in a document. We further show that manipulating the distribution of document lengths and of positional embeddings only marginally mitigates such problems. Our results suggest that even though document-level MT is computationally feasible, it does not yet match the performance of sentence-based MT.

1 Introduction

Statistical and neural machine translation (MT) architectures (Koehn, 2020) have been designed to process isolated sentences, limiting their ability to properly handle discourse phenomena, such as coherence and cohesion, the modelling of which requires longer contexts (Fernandes et al., 2023). A first step to address this shortcoming has been to augment the source and/or the target side with a couple of preceding sentences (Tiedemann and Scherrer, 2017). Multiple approaches to encode and fully exploit such extended contexts have been proposed (Popescu-Belis, 2019; Maruf et al., 2021; Castilho and Knowles, 2024) and have been shown to improve the ability of MT engines to preserve local discourse coherence and cohesiveness through word-sense disambiguation or the resolution of anaphoric references (Bawden et al., 2018; Voita et al., 2018). Most of these approaches continue to process texts on a per-sentence basis with an extended context, even though attempts have also been made to process continuous chunks of texts comprising several sentences (Scherrer et al., 2019; Lopes et al., 2020; Ma et al., 2020, 2021; Lupo et al., 2022a; Wu et al., 2023).

The ability of today's neural MT models—relying on encoder-decoder or decoder-only architectures—to handle large context lengths, up to thousands of tokens (Peng et al., 2024), opens new perspectives to go beyond *context-augmented MT* and develop fully-fledged *document-level MT*, where the entire document context is available at once, and where the target text is generated in a single pass.[1] Two main technical novelties have made this possible: (a) more efficient computation in the attention layers (Tay et al., 2022) and (b) changes in the design of positional encodings (PEs). In particular, replacing the sinusoidal absolute PEs (APEs) of (Vaswani et al., 2017) with methods like ALIBI (Press et al., 2022) and RoPE (Su et al., 2024), which lend themselves well to length extrapolation (Sun et al., 2023; Zhao et al., 2024), seems to make today's transformers amenable to the processing of arbitrarily long contexts (Mohtashami and Jaggi, 2023; Han et al., 2024).

In this work, we challenge the ability of contemporary MT models to effectively handle long spans of texts. For this, we develop a new methodology for assessing the impact of length variations on MT performance. We perform a series of controlled experiments with two representative neural MT systems, where the same documents are processed by chunks of increasing lengths in a document-level manner and show that (a) MT performance tends

[1]These perspectives are, for instance, explored in the latest edition of the WMT shared task on General Machine Translation (Kocmi et al., 2024).

Proceedings of Machine Translation Summit XX Volume 1, pages 4–23, June 23-27, 2025

to degrade with the length of the source document, (b) length issues happen even for in-distribution lengths and get worse when extrapolating to unseen document lengths, and (c) most of the degradation happens in the final parts of the translation. Hypothesising that this may be due to a mismatch between the distribution of train and test PEs, we explore a possible mitigation, which flattens the distribution of PEs during training. We observe a consistent improvement of automatic metric scores for the APE-based vanilla encoder-decoder model, while the RoPE based decoder-only model remains mostly unaffected. In summary, our main contributions are: (a) a new approach to the detection and diagnosis of length issues in document-level MT, (b) a new variant of the SHAPE (Kiyono et al., 2021) method, which improves the distribution of PEs during training, and (c) a confirmation that (perhaps for lack of an appropriate document-level evaluation tool) sentence-level MT remains a strong baseline in most settings.

2 Related work

2.1 Document-level MT

Previous attempts to incorporate more contextual information in MT models can be roughly categorized into two categories: context-augmented MT, also called *Doc2Sent* in (Sun et al., 2022), and document-level MT, also called *Doc2Doc*. Recent surveys of this field include (Popescu-Belis, 2019; Maruf et al., 2021; Castilho and Knowles, 2024).

Translation of discourse phenomena, such as lexical consistency, reference, and word sense disambiguation, requires inter-sentential context (Bawden et al., 2018; Wong et al., 2020; Fernandes et al., 2023). This has motivated the integration of extended (local) contexts in *Doc2Sent* models. Such approaches include concatenation-based methods (Tiedemann and Scherrer, 2017); architecture adaptations to process context in different components of the same encoder (Ma et al., 2020; Wu et al., 2023), in a dedicated encoder (Voita et al., 2018; Zhang et al., 2018), or via hierarchical attention networks (Miculicich et al., 2018; Maruf et al., 2019; Yin et al., 2021); cache-based methods using a short-term MT memory (Maruf and Haffari, 2018; Tu et al., 2018; Yang et al., 2019; Dobreva et al., 2020) and multi-pass decoding algorithms (Voita et al., 2019; Yu et al., 2020; Kang et al., 2020).

Translating sentence by sentence, even with augmented contexts, still fails to capture phenomena related to coherence and consistency (Fernandes et al., 2023), motivating *Doc2Doc* approaches to process documents as a whole. This can be done with concatenation-based methods (Tiedemann and Scherrer, 2017; Sun et al., 2022; Karpinska and Iyyer, 2023), along with sliding window attention (Zhuocheng et al., 2023; Liu et al., 2023) and group attention (Bao et al., 2021) to address the issue of quadratic complexity. Other strategies include focusing on improving training through data augmentation with a balanced length distribution (Sun et al., 2022) and richer context-dependent phenomena (Lupo et al., 2022a; Wu et al., 2024), or on better training strategies with multilingual denoising pre-training (Lee et al., 2022), adapted loss functions (Lupo et al., 2022b), and enriched positional encodings (Li et al., 2023; Lupo et al., 2023). Multiple methods have recently emerged for large language models (LLMs) (Wang et al., 2023), which also show a decline in translation quality as input length increases (Wang et al., 2024).[2] These include a two-stage training recipe with the use of a monolingual corpus and high-quality parallel documents (Xu et al., 2024; Alves et al., 2024), and applying LLMs as post-editors (Koneru et al., 2024).

2.2 Extrapolating PEs

Since self-attention is position-agnostic, PEs are used to provide position information in Transformer models. PEs embed the absolute token position (APEs) (Vaswani et al., 2017), or the relative distance between tokens (RPEs) (Shaw et al., 2018; Raffel et al., 2020; Press et al., 2022), with RoPE (Su et al., 2024) being the go-to approach in recent LLMs such as Llama2 (Touvron et al., 2023). Despite RPEs yielding better length extrapolation ability than APEs, both of them struggle to efficiently extrapolate input lengths beyond the predefined maximum training length (Dai et al., 2019; Chen et al., 2023; Peng et al., 2024; Zhao et al., 2024), motivating the development of input extension methods for PEs.

For APEs, SHAPE (Kiyono et al., 2021) offsets all indices in a sequence by some random values. Its authors show that this simple technique mimics the computation of RPEs at a much smaller cost and helps to improve the interpolation abilities of a vanilla encoder-decoder model, as measured by BLEU (Papineni et al., 2002) with long pseudo-documents. Our experiments confirm that

[2]They also confirmed the effectiveness of training LLMs on documents of varied sizes (similar to Sun et al. (2022)).

this technique is effective using actual document contexts and a sounder experimental methodology, based on paired tests, and using COMET (Rei et al., 2020). Sinha et al.'s (2022) experiments adopt a setting similar to ours, offsetting the absolute value of APEs' input to evaluate their ability to capture relative distances between tokens. Their results, like ours, illustrate the lack of robustness of APEs and suggest that they overfit their training data.

For RPEs, especially RoPE, both position interpolation (PI) and position extrapolation methods have been proposed. PI methods interpolate positions to extrapolate context length directly during inference or through fine-tuning (Chen et al., 2023; Peng et al., 2024). The position extrapolation methods aim to extend context using documents that are shorter than the predefined maximum length. For example, RandPos (Ruoss et al., 2023) randomly maps position indices to a much larger interval with the original word order, and PoSE (Zhu et al., 2024) divides each training sequence into N chunks and adjusts the position indices of every chunk except the first one by adding a uniformly sampled offset, within the scope of a predefined maximal length.

3 Methods and Metrics

3.1 Holistic Document-Level MT

Compared to sentence-based MT, holistic document-Level MT (Doc2Doc) possesses several appealing features, as it gives access to all the available textual context. This should enable the MT system to improve on global aspects pertaining for instance to coherence and cohesion. However, Doc2Doc also introduces several new challenges compared to the Sent2Sent scenario:

1. in Doc2Doc, input texts are longer, causing a computational overhead due to the quadratic complexity of attention (Tay et al., 2022).
2. for longer inputs, attention weights are spread over a larger number of tokens (Herold and Ney, 2023); however, at each decoding step, most attention needs to remain concentrated on the corresponding local source context (Bao et al., 2021). This is in contrast with Doc2Sent, where sentence alignment is readily available.
3. decoding longer sequences increases the impact of search errors and of exposure bias (Ranzato et al., 2016). Beam search also becomes more difficult due to the input length.
4. output sentences may not always stand in one-to-one correspondence with source sentences, which complicates the computation of automatic metrics, which are designed to evaluate one-to-one mappings between hypotheses and references.

These differences motivate our main research questions, which we rephrase as: (a) For existing models, does *Doc2Doc* bring more benefits than disadvantages compared with *Sent2Sent*? (b) How do these results vary with the input document length? (c) Which methods and metrics can we use to automatically evaluate the impact of length differences?

3.2 Shades of BLEU

Answering such questions requires metrics for comparing holistic translations with sentence-based translations: as the number of segments produced by the former may differ from the number of source segments, a basic requirement is that they allow the evaluation of translation hypotheses with more (or fewer) sentences than the source (for quality estimation scores) and/or the reference (for reference-based metrics). However, most existing document-level MT approaches still rely on BLEU (Papineni et al., 2002), despite its well-documented shortcomings (Callison-Burch et al., 2006; Reiter, 2018; Mathur et al., 2020; Dahan et al., 2024); or rather a variant dubbed *d-BLEU* by Liu et al. (2020).[3] We accordingly focus on BLEU in this section, noting that the same questions would need to be addressed with any metric relying on sentence-based surface comparison (e.g., METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), BertSCORE (Zhang et al., 2020), PRISM (Thompson and Post, 2020), COMET (Rei et al., 2020), and many others).[4]

BLEU is computed by counting, sentence by sentence, the number of n-grams (for $n \in [1:4]$) shared by each translation hypothesis and its human reference. These counts are aggregated and turned into frequencies, then averaged (geometrically) at the corpus level. Finally, a length penalty

[3] Hendy et al. (2023) also consider a variant of COMET (Rei et al., 2022) while Zhuocheng et al. (2023) introduce d-ChrF, a document-level version of ChrF (Popović, 2015).

[4] We choose to evaluate using the standard metrics, BLEU and COMET, rather than evaluation approaches specifically designed to test the use of increased context. This choice is motivated by the fact that the score differences we observe reveal a significant degradation in translation quality for longer documents, indicating greater problems than those targeted by finer-grained evaluation techniques.

is applied to degrade the score when the cumulated length of the hypotheses is shorter than that of the references. BLEU is a corpus-level score that depends on sentence alignments. d-BLEU is also a global score but counts common n-grams at the document level. As a consequence, d-BLEU, which records matches for larger spans than BLEU, delivers higher scores, as the opportunities to match n-grams are greater for a wider window.[5] These two scores cannot be compared, and we contend that their shortcomings make them inappropriate for analysing length-related issues in MT.

An alternative to d-BLEU is to perform evaluation at the document level, rather than the corpus level. This can be implemented either as (a) calculating one BLEU score (with realignment) per document, then averaging at the corpus level or (b) calculating the equivalent of sentence-level BLEU scores (Lin and Och, 2004) but where each segment is a concatenated full document rather than a sentence. However, (a) counts matches at the sentence-level, which requires a realignment between translated and reference sentences and may introduce some measurement noise. Therefore, our experiments use method (b) to compute document-level scores, hereafter referred to as **ds-BLEU** scores.

3.3 Evaluating Length Issues in MT

Another recurring methodological caveat with length-related evaluation is related to the way scores are compared. For instance, in (Sun et al., 2022, Figure 1) BLEU scores are reported for buckets of sentences of varying lengths in a plot which suggests that performance increases with length (up to a certain extent). Such vizualisations are misleading, as global BLEU scores should only be compared when measured with the same corpus.

What we propose instead is to compare matching automatic translation scores for a set of inputs $\mathcal{S} = \{s_1 \dots s_T\}$, systematically varying the translation models M in $\{M_1 \dots M_N\}$ and the length of the translation window $W \in \{W_1 \dots W_K\}$. For each pair of settings, we can perform a paired t-test for the average score difference and decide whether two configurations (M_i, W_k) and (M_j, W_l), each associating a system and a length, are statistically different, and if so, which of the two is the best.

[5]This effect is well known, e.g. in (Koehn and Knowles, 2017, Figure 7), where BLEU increases when considering sentence groups of increasing lengths (at least for a certain length range), where we would expect a decrease, as the length is often linked to syntactic complexity and therefore to translation difficulty. We reproduce this observation in Figure 2.

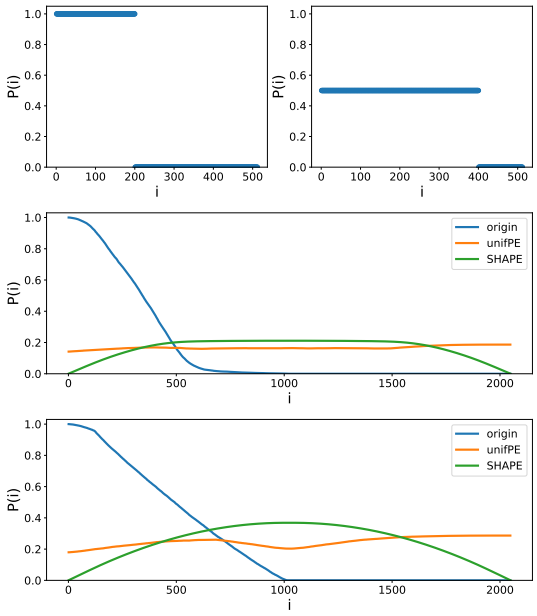


Figure 1: Top: probability of observing training position i ($P(i)$) for a sentence of length $l = 200$, with standard training ($k_i = 0$, left) and with our uniform sampling scheme (right) for $M = 512$. Middle: original, UNIFPE, and SHAPE $P(i)$ for training set **TED-G** and $M = 2048$. Bottom: original, UNIFPE and SHAPE $P(i)$ for **TED-U** and $M = 2048$.

In our experiments, we consider two ways of presenting $\mathcal{S}$: (a) at the document level, where each s_i is a document and the evaluation is the ds-BLEU score introduced in Section 3.2,[6] and (b) at the sentence level, where each s_i is a sentence and the associated metric is COMET (Rei et al., 2020).[7] For (b) we need to realign translation hypotheses with their references. This can be performed with the method of Wicks and Post (2022),[8] or with that of Matusov et al. (2005),[9] which has long been used for evaluating speech translation systems, and which we adopt.[10] Variations in configurations (M, W) are obtained by changing the translation engine and the length of input source texts. In all cases, score comparisons are performed on identical source texts.

[6]We use SacreBLEU (Post, 2018) with signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.0; the parameter *eff* is set to yes for ds-BLEU.

[7]Using the library `https://github.com/Unbabel/COMET` with the default model `wmt22-comet-da`.

[8]Junczys-Dowmunt (2019)'s approach includes a set of tags that constrain input and output to have the same number of sentences, see also (Li et al., 2023).

[9]`https://www-i6.informatik.rwth-aachen.de/web/Software/mwerSegmenter.tar.gz/`.

[10]The per-sentence COMET scores are averaged at the document level to be associated with document lengths, or at the corpus level to assess global translation quality.

The same technique is also used to measure the impact of the position within a document on translation quality. The question we study is whether quality remains constant across a document, or whether it tends to decrease when sentences are processed at higher position indices. For this, we consider groups of sentences translated at varying starting positions with multiple systems and compare the differences between COMET scores with a paired difference test. Details regarding the corpus and window sizes are given in Section 5.1.

4 Manipulating the Distribution of PEs

A basic requirement for document-level systems is that they should be trained, or at least fine-tuned, with long text inputs, ideally with complete documents. Using the empirical document length distribution may however not be ideal, as it yields very skewed distributions of PEs where small position indices are over-represented. We discuss two approaches to obtain more balanced distributions.

4.1 Distribution of PEs

A training sequence of length l yields examples for all indices in $\{1, \ldots, l\}$. For a complete corpus, position index 1 will be observed for all inputs, while the last index of the longest sequence will likely only be observed once. Training with the "natural" distribution of document lengths is therefore likely to overfit to smaller position indices while underfitting to larger ones, hindering the ability to handle long texts or extrapolate to lengths unseen in training (Peng et al., 2024; Zhu et al., 2024).

A first way to improve the distribution of token positions seen in training is to increase the representation of long documents in the training data while keeping a good balance with shorter ones (Bao et al., 2021; Sun et al., 2022). This is easy to do in our controlled setting (see Section 5.1). As our experiments show, this significantly improves automatic scores for the context lengths seen during training. An alternative, which allows us to better study the effect of PE distributions in training, is to directly manipulate the indices (for a fixed length distribution). The UNIFPE algorithm, introduced below, is one way to achieve this.

4.2 Uniform SHAPEs (UNIFPE)

We assume a training set of texts $s_1 \ldots s_N$ of respective lengths $l_1 \ldots l_N$, and a maximum model length of M, with $\forall i, M > l_i$. Training with text s_i creates training samples for positions i in $[1\!:\!l_i]$. For the whole corpus, positions from 1 (observed N times) to $l_{max} = \max_{i=1\ldots N}(l_i)$ are observed, with larger indexes being less trained than smaller ones. Positions indices in $[l_{max} : M]$ are never observed. We wish to make the training PE distribution more even, so that all positions in $[1 : M]$ are equally well-trained, which should also help to extrapolate PEs for indices larger than l_{max}.

This can be achieved by shifting the starting index of every s_i by some offset k_i, making it possible to train with PEs in $[1 + k_i : l_i + k_i]$. How should k_i be chosen? Randomly choosing $k_i = 0$ or $k_i = l_i$ with probability $1/2$ makes the probability of observing any index in $[1\!:\!2l_i]$ equal to $1/2$. This can be generalised to choose k_i with uniform probability $1/m$ among $\{0, l_i, \ldots, (m-1) * l_i\}$, with $m = \lfloor M/l_i \rfloor$. However, doing so implies that indices in $[m * l_i : M]$ are never observed. We compensate for this as follows: before sampling k_i, we modify the set of possible shifts by adding $r_i = M - m * l_i$ to all values larger than a random index $l \in [1 : m]$. In other words, k_i is sampled from $\{j * l_i + r'_{i,j}, j = 0 \ldots m-1\}$, with $r'_{ij} = 0$ if $j < l$ and r_i otherwise. Sampling k_i independently for each text s_i in each training batch ensures that all indices are uniformly represented. A formal description of UNIFPE is given in Algorithm 1. Figure 1 illustrates the difference between always starting at position 1 ($\forall i, k_i = 0$) and using our UNIFPE strategy.

This approach is reminiscent of SHAPE (Kiyono et al., 2021); while SHAPE chooses the offset k_i uniformly at random in a fixed interval to simulate relative PEs, which reduces the frequency of small position indices, we sample k_i non-uniformly to ensure that all indexes are equally represented in training.

5 Experimental Settings

5.1 Datasets

For our experiments, we prepare multiple sets of parallel pseudo-documents based on the EN–FR part of the TEDtalks corpus (Cettolo et al., 2012).

Training and validation sets Our training set consists of pseudo-documents from both the training and validation splits of **IWSLT-2016**.[11] Our goal is to simulate real corpora of parallel documents with source documents shorter than a certain

[11] https://wit3.fbk.eu/2016-01

	TED-full		TED-G		TED-U	
	train	dev	train	dev	train	dev
Count	1831	19	15625	160	10582	106
Length	2915	2861	341	339	504	512

	sent	256	512	768	1024	1200	1600	2048	doc
Count	5103	503	261	184	142	123	100	80	52
Length	23	233	450	638	827	955	1175	1468	2259

Table 1: Left: Statistics of the TED talks training and dev sets. Right: Statistics of the TED talks test sets from IWSLT **tst2014**, **tst2015**, **tst2016** and **tst2017**. 'Count' denotes the number of parallel pseudo-documents, 'Length' denotes the average length of source (i.e. English) pseudo-documents (in NLLB tokens).

	l_{max}	2014	2015	2016	2017
NLLB	sent	45.1 (0.97)	43.9 (0.98)	41.7 (1.00)	41.8 (1.00)
	256	33.9 (0.82)	35.4 (0.84)	33.3 (0.86)	33.5 (0.87)
	512	14.6 (0.44)	16.0 (0.56)	15.2 (0.52)	13.8 (0.49)
	768	7.3 (0.27)	7.9 (0.32)	10.0 (0.46)	6.7 (0.27)
	1024	8.8 (0.56)	7.4 (0.51)	7.5 (0.50)	6.5 (0.48)
TOWERBASE	sent	43.4 (0.98)	42.9 (0.99)	39.7 (1.00)	38.7 (1.00)
	256	44.0 (0.96)	42.8 (0.98)	40.9 (1.00)	39.4 (1.00)
	512	42.9 (0.96)	39.8 (0.98)	39.9 (1.00)	40.6 (1.00)
	768	39.6 (0.98)	39.0 (0.97)	38.1 (0.99)	39.9 (1.00)
	1024	38.5 (0.98)	33.1 (0.99)	35.4 (1.00)	35.4 (0.98)
	1200	37.4 (0.92)	35.5 (0.98)	36.2 (1.00)	35.6 (0.98)
	1600	33.3 (0.96)	34.9 (0.96)	26.7 (0.94)	31.0 (0.97)
	2048	24.0 (0.97)	27.7 (0.95)	27.2 (0.96)	23.5 (0.87)

Table 2: ds-BLEU scores (and brevity penalty) for NLLB200-DISTILLED-600M and TOWERBASE-7B.

length l_{max} – using $l_{max} = 1024$. We split all document pairs whose source side is longer than 1024 tokens into fragments.[12] For each document pair, we iterate the following procedure: (1) sample a maximum pseudo-document length l'_i following the same Gaussian-like length distribution as the full TED talks with $l'_i < l_{max}$, (2) concatenate consecutive sentence pairs up to l'_i to form a training pseudo-document s_i. The resulting distribution of document lengths is displayed in Figure 3 in Appendix A.3. The development set is built similarly, using document pairs from IWSLT **tst2010** and **tst2011**. We denote these training datasets as **TED-G** (G for Gaussian). As discussed in Section 4, we consider another dataset generation strategy, which produces a more balanced length distribution, for which we do as above but we sample uniformly: $l'_i \sim U(128, l_{max})$.[13] Fine-tuning with the resulting **TED-U** corpus allows us to contrast two distributions with differences in document length.

Test sets To evaluate MT systems for their ability to handle documents of varying sizes and extrapolate beyond the training samples, we build a series of test sets of increasing document lengths. For each document in IWSLT **tst2014**, **tst2015**, **tst2016** and **tst2017**, we accumulate consecutive sentence pairs into parallel pseudo-documents such that all resulting source texts have a length close to l_{max}, with $l_{max} \in \{256, 512, 1024, 1200, 1600, 2048\}$.[14] Contrarily to training sets, test sets are homogeneous in length. Statistics are in Table 1 with more details in Appendix A.3. Evaluation is always performed with complete original talks, after concatenating and aligning all the corresponding parts.

5.2 Models

We used the UNIFPE algorithm to fine-tune two pre-trained MT systems that were not trained with TED talks. As UNIFPE is designed for APEs, we considered NLLB200-DISTILLED-600M[15] or NLLB for short (Costa-jussà et al., 2024) as a representative encoder-decoder model based on APEs. NLLB is a 12-layer encoder-decoder multilingual MT model pre-trained on 200 languages. We used the HuggingFace implementation, which relies on sinusoidal APEs (Vaswani et al., 2017). We also perform fine-tuning with SHAPE for comparison. We refer to the specific MT systems with respect to their fine-tuning method (FT, UNIFPE or SHAPE), backbone model (e.g. NLLB) and training corpus (U for **TED-U** or G for **TED-G**. More precisely, we denote MT systems trained on **TED-U** (resp. **TED-G**) as FT-NLLB-U (resp. FT-NLLB-G), UNIF-NLLB-U (resp. UNIF-NLLB-G) when fine-tuning with UNIFPE, and SHAPE-NLLB-U (resp. SHAPE-NLLB-G).

We also experiment with an LLM-based architecture, TOWERBASE-7B[16] (Alves et al., 2024) (TOWERBASE for short), derived from Llama2 (Touvron et al., 2023) using translation-related

[12] All statistics counted in tokens use the tokeniser of NLLB (Costa-jussà et al., 2024).

[13] Short pseudo-documents continue to be slightly over-represented, because the last pseudo-document in any given talk is often strictly shorter than the desired length l'_i.

[14] At the end of each talk, we concatenate the last parallel sentences into the last pseudo-document if they are shorter than 50 to avoid exceedingly short parallel sequences.

[15] https://huggingface.co/facebook/nllb-200-distilled-600M

[16] https://huggingface.co/Unbabel/TowerBase-7B-v0.1

	NLLB	FT-U	Unif-U	SHAPE-U	FT-G	Unif-G	SHAPE-G
sent-256	9.2	0.8	-2.1	-2.5	0.4	-0.7	-1.7
256-512	19.1	-	-0.4	1.4	-	-	2.1
512-768	6.9	-	-0.6	-	5.9	2.5	5.3
768-1024	-	0.5	-	-	7.2	3.0	3.7
1024-1200	2.2	3.5	1.9	4.1	4.0	3.3	4.1
1200-1600	-	6.8	6.5	5.4	5.8	5.5	4.9
1600-2048	1.9	5.2	4.5	3.1	4.2	5.7	6.0
sent-256	16.7	3.5	1.7	1.3	2.7	2.1	1.9
256-512	20.7	-	-0.4	2.4	0.6	-	4.2
512-768	5.6	-	-	-	11.6	7.2	8.1
768-1024	5.2	2.3	0.8	-	10.4	7.6	7.8
1024-1200	-	7.4	4.3	6.9	3.8	4.7	4.6
1200-1600	6.1	9.6	13.4	8.3	5.4	5.5	5.6
1600-2048	-	5.1	5.0	5.9	3.9	5.6	5.0

	TOWER	FT-U	Unif-U	SHAPE-U	FT-G	Unif-G	SHAPE-G
sent-256	-	-	-	-	-	-	-
256-512	-	0.9	0.8	0.8	0.6	0.6	0.5
512-768	-	-	-	-	0.6	-	-
768-1024	3.4	-	1.0	1.2	1.7	1.2	2.1
1024-1200	-	-	-	-	-	-	-
1200-1600	4.7	1.7	2.1	-	1.6	2.0	1.5
1600-2048	5.9	7.5	6.5	7.3	8.1	7.8	7.3
sent-256	3.9	2.3	2.4	2.3	2.3	2.3	2.2
256-512	-	0.4	-	-	0.2	0.3	0.3
512-768	-	-	-	0.5	0.3	-	-
768-1024	2.9	1.0	0.5	-	1.1	0.8	1.2
1024-1200	-	-	1.0	0.9	-	-	-
1200-1600	6.2	1.7	1.8	-	-	1.9	1.8
1600-2048	8.7	10.0	8.9	9.1	11.0	10.2	9.2

Table 3: Average differences evaluated on **ds-BLEU** (top) of full TED talks and on 100×**COMEST** (bottom) of realigned parallel sentences, between translations in increasing context size, for NLLB (left) and TOWERBASE (right) models. U and G respectively denote **TED-U** and **TED-G**. A positive value means that shorter segments result in higher scores than longer ones. Text in olive for p-values > 0.01. - for p-values > 0.05.

tasks. TOWERBASE uses RoPE (Su et al., 2024) to encode RPEs. As mentioned by Peng et al. (2024), they nonetheless encode some form of APE signal in some dimensions, and may therefore be also mildly impacted by the PE training distribution. We refer to the models based on TowerBase as FT-TOWER-U (resp. FT-TOWER-G), UNIF-TOWER-U (resp. UNIF-TOWER-G) and SHAPE-TOWER-U (resp. SHAPE-TOWER-G) the model fine-tuned on **TED-U** (resp. **TED-G**) with original PEs, UNIFPE or SHAPE.

Both backbone models were pretrained with large amounts of EN–FR data; we focus exclusively on the EN into FR direction. Details on fine-tuning and decoding parameters can be found in Appendix A.4.

6 Results and Analyses

6.1 Length Issues

We report the ds-BLEU (Table 2) and COMET (Appendix, Table 7) scores of the pretrained models NLLB and TOWERBASE for multiple test sets, varying the average input segment lengths from one sentence to the maximum input length used in training.[17] For NLLB, we observe a drop of around 10 ds-BLEU points and about 0.2 COMET points when translating test sets of $l_{max} = 256$ instead of isolated sentences. Scores and their associated brevity penalties (BPs) only get worse with larger context lengths. For TOWERBASE, the decrease in BLEU is more progressive, with a sharp decline for all test sets for $l_{max} > 1024$. The related COMET scores plummet immediately with a context size of 256. Even though TOWERBASE is based on Llama2, which accepts inputs up to 4096 tokens, the continued pretraining that was used mostly uses isolated sentences, which introduces an inductive bias affecting its ability to translate long texts.

As expected, document-level fine-tuning (DLFT) has a strong positive impact (see Appendix, Table 11). However, the length issues remain.

Length Bias We performed paired comparisons for the translation of our test sets with increasing text lengths for each MT system as presented in Section 3.3. Results are given in Table 3, where a positive difference (e.g. 9.2 for NLLB in line "sent-256") means that the translation of shorter segments (here: sentences) yield better scores than that of longer ones (256 tokens). Scores in the same column are comparable. Except for a handful of configurations, translating longer texts is never better than translating short ones. We conclude that in our experimental settings, the disadvantages associated with long inputs (Section 3.1) overwhelm the benefits of a complete context. These length issues result in large score degradations and are not easily fixed by simple manipulation of PEs. We also observe that results obtained with COMET and ds-BLEU sometimes disagree. These cases are rare, though, suggesting that our results are robust.

Document-level Tuning with UNIFPE Again using the paired comparison methodology, we compare the performance of DLFT with original PEs, UNIFPE and SHAPE. As shown in the left and middle parts of Table 4, fine-tuning using UNIFPE leads to steady improvements in translation scores for all test lengths, especially for systems fine-

[17] As explained in Section 3.2, these COMET scores require the realignment of target sentences with the reference.

	TED-U		TED-G		FT	Unif	SHAPE
	FT vs Unif	FT vs SHAPE	FT vs Unif	FT vs SHAPE	U vs G	U vs G	U vs G
sent	3.3 (0.00)	4.0 (0.00)	1.2 (0.00)	2.4 (0.00)	-	-2.1 (0.00)	-1.6 (0.00)
256	-	0.7 (0.00)	-	-	-	-0.6 (0.01)	-0.7 (0.00)
512	-0.5 (0.01)	1.7 (0.01)	-	2.3 (0.00)	-0.7 (0.00)	**-0.4 (0.01)**	-
768	-0.8 (0.00)	2.7 (0.00)	-3.7 (0.00)	-	5.5 (0.00)	2.6 (0.00)	4.5 (0.00)
1024	-	1.6 (0.00)	-7.8 (0.00)	-1.8 (0.04)	12.2 (0.00)	5.1 (0.00)	8.8 (0.00)
1200	-2.3 (0.00)	2.3 (0.02)	-8.5 (0.00)	-	12.7 (0.00)	6.5 (0.00)	8.8 (0.00)
1600	**-2.6 (0.01)**	-	-8.8 (0.00)	**-2.6 (0.01)**	11.8 (0.00)	5.6 (0.00)	8.3 (0.00)
2048	**-3.3 (0.00)**	-	-7.3 (0.00)	-	10.7 (0.00)	6.8 (0.00)	11.3 (0.00)
sent	1.9 (0.00)	2.9 (0.00)	0.9 (0.00)	1.4 (0.00)	-	-0.9 (0.00)	-1.4 (0.00)
256	-	0.8 (0.00)	**0.4 (0.00)**	**0.7 (0.00)**	**-0.8 (0.00)**	-0.5 (0.01)	-0.9 (0.00)
512	-0.6 (0.02)	2.9 (0.00)	-	4.3 (0.00)	-0.4 (0.03)	-	-
768	-0.7 (0.00)	4.7 (0.00)	-4.7 (0.00)	-	11.2 (0.00)	7.2 (0.00)	7.3 (0.00)
1024	**-2.2 (0.00)**	3.3 (0.00)	-7.5 (0.00)	-1.8 (0.02)	19.3 (0.00)	14.0 (0.00)	14.2 (0.00)
1200	-5.3 (0.00)	2.7 (0.02)	-6.5 (0.00)	-	15.7 (0.00)	14.5 (0.00)	11.9 (0.00)
1600	-	-	-6.4 (0.00)	-	11.5 (0.00)	6.6 (0.00)	9.2 (0.00)
2048	-	**2.1 (0.04)**	-4.7 (0.00)	-	10.4 (0.00)	7.2 (0.00)	8.3 (0.00)

Table 4: Average difference (and p-values) in **ds-BLEU** (top) evaluated on full TED talks and 100×**COMET** (bottom) evaluated on realigned sentences for NLLB. Left and middle: paired comparison between fine-tuning with the original PEs (FT), UNIFPE (Unif) and SHAPE on **TED-U** and **TED-G** respectively. Right: differences between fine-tuning on **TED-U** (U) and **TED-G** (G). - for p-values > 0.05. Bold values when the two metrics disagree on significativity.

	NLLB	FT-U	Unif-U	SHAPE-U	FT-G	Unif-G	SHAPE-G
p_0-p_1	12.7	-	-	4.0	1.6	3.0	5.0
p_1-p_2	7.2	-	-	-2.0	1.9	2.4	-
p_2-p_3	2.9	1.0	-1.3	-	2.4	-	-2.1
p_3-p_4	7.2	5.5	4.6	7.3	26.1	10.7	13.9
p_4-p_5	-	3.9	-	-	8.3	4.7	4.3
p_5-p_6	3.3	31.6	27.1	19.5	6.1	15.4	15.3

	TOWER	FT-U	Unif-U	SHAPE-U	FT-G	Unif-G	SHAPE-G
p_0-p_1	1.2	-	-	-	-	-	-
p_1-p_2	-	-	-	-	0.6	0.6	-
p_2-p_3	-	-	-	-	-	-	-
p_3-p_4	4.9	1.1	0.6	1.4	1.1	1.3	1.1
p_4-p_5	1.7	1.4	2.1	1.7	2.0	1.8	2.2
p_5-p_6	26.3	24.5	26.0	25.2	27.1	26.4	27.0

Table 5: Average difference of 100×COMET-score evaluated on 794 sentence pairs, translated at different positions (e.g. p_0 and p_1 with $p_0 < p_1$) by NLLB-based systems (top) and TOWERBASE-based systems (bottom). Olive text for p-values > 0.01. - for p-values > 0.05.

tuned with the unbalanced corpus (**TED-G**). The only exception is for sentence-level translations, which remain marginally better using standard DLFT than with UNIFPE. In contrast, SHAPE only improves DLFT performance on the **TED-G** corpus and for translation windows greater than 1024 tokens, due to the under-representation of small position indices during training, as shown in Figure 1. As Tables 11 to 15 show, these improvements remain moderate, and the length issues continue to strongly impact translation scores, especially for test documents of 1024 tokens or more. For TOWERBASE, UNIFPE does not yield any significant difference with standard DLFT, and SHAPE occasionally delivers slight improvements (see Appendix, Table 16), likely because this model relies on RPEs. From these comparisons, we conclude that UNIFPE partly resolves length issues for NLLB, but hardly changes the situation for TOWERBASE.

Impact of Data Distribution In the right part of Table 4, we evaluate the impact of the length distribution during fine-tuning for NLLB: the balanced distribution (**TED-U**) slightly but consistently underperforms the use of **TED-G** for short documents (fewer than 512 tokens), a trend that is reversed for longer documents with strong improvement (over 768 tokens). Manipulating the distribution of PEs with UNIFPE reduces the gap between the two fine-tuning corpora and makes the model more robust to document lengths rarely observed (or even unobserved) during fine-tuning. This analysis again reveals small differences between using ds-BLEU and COMET scores: in nine cases out of 56 comparisons (marked in bold), one metric detects a difference that is non-significant for the other.

6.2 Position Bias

To investigate potential translation issues related to large position indices, we collected the 794 sentences that come from the final part of long talks and for which varying the window length also varied the position index. For each of them, we have seven translations, corresponding to positions $\{p_0^j, \ldots, p_6^j\}, j \in \{1, \ldots, 794\}$. The av-

	256	512	768	1024	1200	1600	2048
NLLB	0.04	0.35	0.49	0.66	0.64	0.74	0.81
FT-U	0.01	0.03	0.08	0.09	0.13	0.26	0.44
Unif-U	0.01	0.03	0.07	0.10	0.20	0.34	0.32
SHAPE-U	0.03	0.08	0.11	0.20	0.36	0.39	0.46
FT-G	0.01	0.03	0.11	0.31	0.40	0.57	0.69
Unif-G	0.01	0.04	0.16	0.20	0.27	0.25	0.36
SHAPE-G	0.02	0.05	0.12	0.17	0.21	0.24	0.28

	256	512	768	1024	1200	1600	2048
TOWER	0.01	0.05	0.13	0.30	0.29	0.45	0.64
FT-U	0.01	0.05	0.10	0.15	0.13	0.23	0.59
Unif-U	0.02	0.05	0.09	0.15	0.16	0.28	0.61
SHAPE-U	0.02	0.05	0.08	0.15	0.17	0.21	0.59
FT-G	0.02	0.05	0.10	0.15	0.19	0.26	0.62
Unif-G	0.02	0.07	0.10	0.16	0.20	0.28	0.64
SHAPE-G	0.02	0.05	0.07	0.17	0.17	0.27	0.62

Table 6: Percentage of pseudo-documents among IWSLT **tst2014-2017** in which 10-gram repetition is detected in the translation given by NLLB-based (top) and TOWERBASE-based models (bottom).

erage values for $\{p_0^j, \ldots, p_6^j\}$ are $\{p_0, \ldots, p_6\} = [66, 173, 262, 335, 585, 779, 1477]$. For this subset of sentences, we performed a paired t-test to compare the impact of the position on the translation score (using COMET as the only metric). We observe in Table 5 that in almost all comparisons but three, a small position index is preferable to a larger one. This suggests that one of the main challenges faced by *Doc2Doc* with large context lengths is to control the quality degradation for the final parts of the input text. Here again, UNIFPE slightly mitigates the problem for NLLB models compared with original PEs and SHAPE, but no such improvement is observed for TOWERBASE.

6.3 Repeated n-grams in Translation

One obvious problem with holistic translations produced by NLLB is the generation of outputs that are too short. A closer look at translation outputs also reveals that outputs contain many instances of repeated texts, usually occurring in the final part of the translation. To quantify this problem, we compute the percentage of translations of pseudo-documents in which the repetition of a long n-gram (with $n \geq 10$) is detected. Detailed results are given in Table 6. For all systems and fine-tuning strategies, the percentage of repetitions increases with the length, a problem that seems (for large text lengths) slightly more severe for TOWERBASE, which has a much better BP, than NLLB.

7 Conclusion

In this work, we have studied the ability of current MT architectures to handle long input texts, ideally entire documents, and to translate them holistically. Our analyses are based on systematic comparisons of translation outputs computed with varying input lengths, which are then evaluated with two automatic metrics. They consistently show that, even when the test document lengths match that of the training set and remain within the model limits, the translation scores tend to decrease with the source length, a degradation that mostly impacts sentences occurring far from the beginning of the document. We also show that manipulating the training distribution of lengths or PEs has a positive effect for APE-based models, which vanishes in RoPE-based models like TOWERBASE. These results finally confirm the robustness of sentence-level baselines. They hint at the need to improve existing models to truly benefit from the potential of document-level MT, for instance by constraining the attention mechanism to simulate a form of sentence alignment, by improving the memorization capacities of existing architectures, or by ensuring that the generation algorithm does not eventually get trapped in repetition loops. These are some of the directions we wish to explore in future work.

8 Limitations

The empirical observations reported in this paper are based on one single language direction, and one domain (TEDtalks). This experimental design is motivated by (a) the fact that French-English is considered an easy pair for MT, with large sets of parallel training data; (b) TEDtalks data are a standard benchmark for document-level MT, and crucially contain very long parallel documents, allowing us to implement our evaluation methodology on a large range of length values. Furthermore, these datasets are not included in the training data of our models. We contend that the length issues observed in these favorable conditions for two representative systems would only be worse for more difficult or less-resourced language pairs.

9 Ethics Statement

This study has been performed with standard benchmarks and open-weight models. We do not see any ethical problems with this work.

10 Carbon Impact Statement

The experiments were conducted on a private infrastructure using a single A100 SXM4 GPU, with a carbon efficiency of 0.432 kgCO_2eq/kWh. The average time required for fine-tuning and checkpoint selection was 14.21 hours for the six NLLB models, and 5.6 hours for the TOWERBASE models. The average emissions are estimated to be 2.45 kgCO_2eq for NLLB-based models and 0.97 kgCO_2eq for models derived from TOWERBASE, with no offset applied. These estimations were based on the Machine Learning Impact calculator[18] (Lacoste et al., 2019).

Acknowledgments

This work was supported by the French national agency ANR as part of the MaTOS project.[19] 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 are grateful to the anonymous reviewers for their insightful comments and suggestions and to Paul Lerner for his review and feedback on a preliminary draft of this work.

References

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*.

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*, pages 65–72, Ann Arbor, Michigan.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level 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 3442–3455, Online. 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.

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*, Trento, 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.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *Preprint*, arXiv:2306.15595.

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, Jeff Wang, and NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846. ISBN: 1476-4687.

Nicolas Dahan, Rachel Bawden, and François Yvon. 2024. Survey of Automatic Metrics for Evaluating Machine Translation at the Document Level. Technical report, Inria Paris; ISIR-CNRS .

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Radina Dobreva, Jie Zhou, and Rachel Bawden. 2020. Document sub-structure in neural machine translation. In *Proceedings of the 12th Language Resources*

[18] https://mlco2.github.io/impact#compute

[19] http://anr-matos.fr/

and Evaluation Conference, pages 3657–3667, Marseille, France. European Language Resources Association.

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.

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. 2024. LM-infinite: Zero-shot extreme length generalization for large language models. 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 3991–4008, Mexico City, Mexico. 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.

Christian Herold and Hermann Ney. 2023. Improving long context document-level machine translation. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 112–125, Toronto, Canada. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

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.

Shun Kiyono, Sosuke Kobayashi, Jun Suzuki, and Kentaro Inui. 2021. SHAPE: Shifted absolute position embedding for transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3309–3321, Online and Punta Cana, Dominican Republic. 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 Eighth Conference on Machine Translation*, Miami, Florida, USA. Association for Computational Linguistics.

P. Koehn. 2020. *Neural Machine Translation*. Cambridge University Press.

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.

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.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Chia-Hsuan Lee, Aditya Siddhant, Viresh Ratnakar, and Melvin Johnson. 2022. DOCmT5: Document-level pretraining of multilingual language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 425–437, Seattle, United States. 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.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

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.

Zihan Liu, Zewei Sun, Shanbo Cheng, Shujian Huang, and Mingxuan Wang. 2023. Only 5% attention is all you need: Efficient long-range document-level neural machine translation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 733–743, Nusa Dua, Bali. Association for Computational Linguistics.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

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.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2023. Encoding sentence position in context-aware neural machine translation with concatenation. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 33–44, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Zhiyi Ma, Sergey Edunov, and Michael Auli. 2021. A comparison of approaches to document-level machine translation. arXiv preprint arXiv:1910.07481.

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)*, pages 1275–1284, Melbourne, Australia. Association for Computational Linguistics.

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).

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.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *Proceedings of the Second International Workshop on Spoken Language Translation*, Pittsburgh, Pennsylvania, USA.

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.

Amirkeivan Mohtashami and Martin Jaggi. 2023. Landmark attention: Random-access infinite context length for transformers. In *Workshop on Efficient Systems for Foundation Models @ ICML2023*.

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.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Andrei Popescu-Belis. 2019. Context in neural machine translation: A review of models and evaluations. *Preprint*, arXiv:1901.09115.

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, Belgium, Brussels. Association for Computational Linguistics.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

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. *J. Mach. Learn. Res.*, 21(1).

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016*, San Juan, Puerto Rico.

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.

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

Anian Ruoss, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Róbert Csordás, Mehdi Bennani, Shane Legg, and Joel Veness. 2023. Randomized positional encodings boost length generalization of transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1889–1903, Toronto, Canada. Association for Computational Linguistics.

Yves Scherrer, Jörg Tiedemann, and Sharid Loáiciga. 2019. Analysing concatenation approaches to document-level NMT in two different domains. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 51–61, Hong Kong, China. Association for Computational Linguistics.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. 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 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Koustuv Sinha, Amirhossein Kazemnejad, Siva Reddy, Joelle Pineau, Dieuwke Hupkes, and Adina Williams. 2022. The curious case of absolute position embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4449–4472, Abu Dhabi, United Arab Emirates. 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 seventh conference of the Association for Machine Translation in the America (AMTA)*, pages 223–231, Boston, Massachusetts, USA.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. 2023. A length-extrapolatable transformer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14590–14604, Toronto, Canada. 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. 2022. Efficient Transformers: A Survey. *ACM Computing Surveys*, 55(6):1–28.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

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.

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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Moly-

bog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

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 30*, pages 5998–6008. Curran Associates, Inc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. 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, 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)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

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.

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.

Rachel Wicks and Matt Post. 2022. Does sentence segmentation matter for machine translation? In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 843–854, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

KayYen Wong, Sameen Maruf, and Gholamreza Haffari. 2020. Contextual neural machine translation improves translation of cataphoric pronouns. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5971–5978, Online. Association for Computational Linguistics.

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. Document flattening: Beyond concatenating context for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 448–462, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

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*.

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.

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F.T. Martins, and Graham Neubig. 2021. Do context-aware translation models pay the right attention? 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 788–801, Online. 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.

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*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

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

Liang Zhao, Xiachong Feng, Xiaocheng Feng, Weihong Zhong, Dongliang Xu, Qing Yang, Hongtao Liu, Bing Qin, and Ting Liu. 2024. Length extrapolation of transformers: A survey from the perspective of positional encoding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9959–9977, Miami, Florida, USA. Association for Computational Linguistics.

Dawei Zhu, Nan Yang, Liang Wang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. PoSE: Efficient context window extension of LLMs via positional skip-wise training. In *The Twelfth International Conference on Learning Representations*.

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.

A Appendix

A.1 The UNIFPE Algorithm

The UNIFPE algorithm briefly described in Section 4.2 is formalised in Algorithm 1.

Data: l_i: The input length
Data: M: The target max context length
Data: $\text{List}_{\text{p}_\text{k}}$: the distribution of p_k for each offset k in $[0, M - l_i]$
$\text{List}_{\text{p}_\text{k}} \longleftarrow$ Initialized to 0 for each element
$m \longleftarrow \lfloor M/l_i \rfloor$ nb. of possible non-zero p_k
$R_n \longleftarrow$ the remainder of M divided by l_i
$p_0 \longleftarrow \frac{1}{m}$the probability of each non-zero p_k
if $M < 2l_i$ **then**
  $\text{List}_{\text{p}_\text{k}} \longleftarrow p(k' = 0) = 1$ i.e. $p_{k=0} = 1$
else
  $k^* \longleftarrow$ a random integer in $[0, m)$
  for $k \in [0, M - l_i]$ **do**
    if $k\%l_i == 0$ and $k < k^*$ **then**
      $\text{List}_{\text{p}_\text{k}} \longleftarrow p(k' = k) = p_0$
    end
    if $(k - R_n)\%l_i == 0$ and $k^* < k \leq M - l_i$ **then**
      $\text{List}_{\text{p}_\text{k}} \longleftarrow p(k' = k) = p_0$
    end
  end
end
return $\text{List}_{\text{p}_\text{k}}$

Algorithm 1: UNIFPE: the pseudo-uniform position indices mapping algorithm.

A.2 A Call for Correctly Using BLEU Scores

As illustrated in Figure 2, d-BLEU and ds-BLEU are always larger than BLEU. When BLEU decreases due to the degradation of translation quality, d-BLEU remains stable because of the higher probability to find n-gram matches in longer sequences. In contrast, ds-BLEU consistently decreases when BLEU diminishes, as it applies a macro-average to compute the corpus-level score, which is more sensitive to the translation quality of each document than d-BLEU. Therefore, d-BLEU, ds-BLEU and BLEU are not comparable and d-BLEU is not suitable for analysing length issues in document-level evaluation of MT.

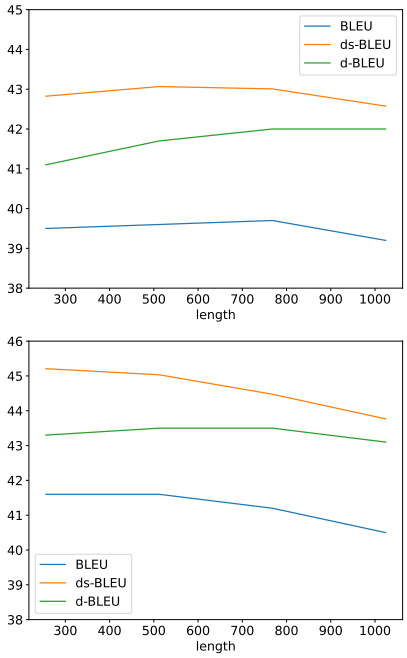


Figure 2: BLEU, ds-BLEU and d-BLEU scores for IWSLT **tst2015**, translating and evaluating *pseudo-documents* of increasing lengths $[256, 512, 768, 1204]$, using FT-NLLB-U (top) and UNIF-TOWER-U (bottom). Note that d-BLEU is computed for pseudo-documents while ds-BLEU is computed for concatenated full talks.

A.3 Data Statistics and Other Details

Full data statistics are given in Tables 8 and 9. All the full TED talks in our corpora start with the title, then the description and the talk before being split into pseudo-documents. `<description>` and `<title>` tags are removed. When preparing our training and validation sets **TED-U** and **TED-G** (see Section 5.1), if concatenating the last sentence pair (x_n, y_n) into the current pseudo-document pair exceeds l_{max}, (x_n, y_n) will yield a single parallel sequence, to respect the maximum length l_{max}. The length distribution is illustrated in Figure 3.

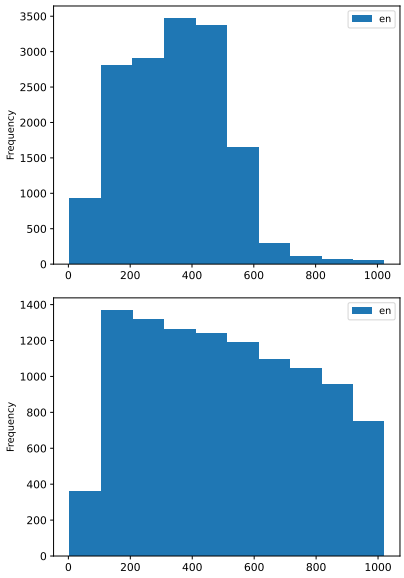


Figure 3: Source document length distribution in the training set of **TED-G** (top) and **TED-U** (bottom).

		2014	2015	2016	2017
NLLB	sent	84	85	85	84
	256	68	69	68	66
	512	49	47	47	46
	768	43	42	40	41
	1024	36	37	36	36
TOWERBASE	sent	84	85	85	85
	256	80	81	82	80
	512	79	80	82	80
	768	78	78	80	80
	1024	76	73	78	76
	1200	73	74	77	76
	1600	70	72	65	68
	2048	52	63	65	57

Table 7: 100×COMET scores for NLLB (top) and TOWERBASE (bottom).

A.4 Experimental Settings

This section presents detailed experiment settings for fine-tuning NLLB and TOWERBASE.

For NLLB, we fine-tuned the pretrained model with learning rate 5e−4, 500 warm-up steps, 4 parallel pseudo-documents per batch and 32 gradient accumulation steps. An early stopping criterion with a patience of 5 epochs is also applied, according to the d-BLEU evaluated on the validation set. For inference on test sets, the beam size is set to 5 and the batch size is set to 4.

For TOWERBASE, We performed supervised fine-tuning using QLoRA (Dettmers et al., 2023) and bfloat16.[20] The batch size is 8 with 2 gradient accumulation steps. The learning rate is 2e−5 adjusted by a *cosine* schedule, without warm-up steps nor packing. We fine-tuned the model for two epochs and saved checkpoints every 50 steps in the second epoch. We then chose the checkpoint with the best d-BLEU on the validation set. Inference is performed without additional in-context examples, with bfloat16 and greedy search.

[20]The prompt for fine-tuning is "Translate the following text from English into French.\nEnglish: SRC\nFrench: TGT", and the zero-shot prompt for the pretrained model TOWERBASE is "English: SRC\nFrench:".

		count	mean	min	max
TED-full	train	1831	2915 /3515	56 /62	8706 /9706
	dev	19	2861 /3460	680 /867	6076 /7590
TED-G	train	15625	341 /411	3 /2	1022 /1460
	dev	160	339 /410	12 /17	959 /1203
TED-U	train	10582	504 /608	3 /1	1020 /1527
	dev	106	512 /620	42 /41	991 /1276

Table 8: Statistics of the TED talks training and dev sets. 'count' denotes the number of parallel pseudo-documents. 'mean', 'min' and 'max' represent the average, minimum and maximum lengths of English/French pseudo-documents respectively, in NLLB tokens.

A.5 Detailed Evaluation Results

The paired comparison and the complete BLEU and COMET scores for each test set are given in Tables 11 to 15.

Document-level Fine-tuning Table 11 reports average differences of automatic scores between fine-tuned MT systems and the corresponding pretrained models (NLLB or TOWERBASE), for varying test document lengths. ds-BLEUs are averaged over 52 complete TED talks and COMET scores are averaged over 5,103 sentences. Fine-tuning significantly improves over base conditions for all lengths, with larger increases for longer test texts, where the baseline scores were initially very poor. Both metrics yield consistent conclusions, except for the sentence-level assessment of NLLB fine-tuned on **TED-U**, which is slightly worse than the baseline according to ds-BLEU (-0.8), but for which COMET detects no difference. For TOWERBASE, DLFT is always beneficial.

Realignment Issues Since the COMET score is sentence-based, its computation requires a realignment between hypotheses and reference sentences in the Doc2Doc scenario. However, due to issues with long documents, translation hypotheses can be incomplete, resulting in empty alignment for some sentences. These untranslated sentences often occur in the final part of long documents. Table 10

	l_{max}	2014 count	2014 min	2014 max	2014 mean	2015 count	2015 min	2015 max	2015 mean	2016 count	2016 min	2016 max	2016 mean	2017 count	2017 min	2017 max	2017 mean
EN	sent	1335	2	112	23	1104	2	119	23	1185	1	151	24	1479	2	162	23
	256	129	65	286	234	107	71	325	234	123	61	255	232	144	65	271	234
	512	68	85	511	443	56	53	510	447	63	56	511	454	74	73	511	456
	768	48	116	767	628	40	86	766	626	45	104	767	635	51	57	767	662
	1024	37	83	1022	815	30	68	1023	835	35	115	1023	817	40	65	1023	844
	1200	32	54	1218	942	26	71	1198	963	31	73	1216	922	34	125	1203	992
	1600	26	135	1597	1160	24	114	1599	1043	23	191	1616	1243	27	229	1635	1250
	2048	20	569	2091	1507	16	176	2072	1565	21	247	2046	1361	23	65	2045	1467
	doc	15	995	4116	2010	12	1256	3359	2086	13	842	3366	2199	12	1909	3722	2812
FR	sent	1335	2	158	28	1104	2	145	27	1185	1	180	29	1479	2	211	27
	256	129	80	380	295	107	85	355	282	123	69	345	276	144	78	375	275
	512	68	106	717	559	56	62	679	540	63	70	672	539	74	83	737	535
	768	48	142	1065	792	40	102	1009	755	45	112	985	755	51	68	1083	776
	1024	37	100	1436	1027	30	64	1349	1007	35	125	1314	970	40	80	1431	990
	1200	32	61	1641	1188	26	85	1577	1162	31	93	1511	1096	34	134	1714	1164
	1600	26	173	2188	1462	24	156	2116	1259	23	209	2074	1477	27	279	2261	1466
	2048	20	657	2613	1901	16	218	2602	1889	21	280	2626	1617	23	80	2714	1721
	doc	15	1289	4983	2534	12	1609	4013	2518	13	1004	4179	2613	12	2473	4464	3299

Table 9: Statistics of the test sets based on talks from IWSLT **tst2014**, **tst2015**, **tst2016** and **tst2017** (see Section 5.1). 'count' refers to the number of parallel pseudo-documents. 'mean', 'min' and 'max' denote the average, minimum and maximum lengths of the source (i.e. English, top) or the reference (i.e. French, bottom) pseudo-documents. All lengths are counted in NLLB tokens.

displays the statistics of empty alignments across all the $5,103$ sentences. This issue is more severe for NLLB models than TOWERBASE models, which is consistent with the poor BP reported in Tables 12 and 13.

	NLLB	FT-U	Unif-U	FT-G	Unif-G
sent	0	0	0	0	0
256	557	6	6	6	6
512	1231	5	9	12	11
768	1618	6	10	250	280
1024	886	53	34	491	486
1200	1179	437	207	576	675
1600	352	465	657	789	840
2048	456	644	843	801	1089

	TOWER	FT-U	Unif-U	FT-G	Unif-G
sent	0	0	0	0	0
256	79	3	8	2	3
512	58	2	2	3	2
768	65	4	3	3	5
1024	45	17	21	19	22
1200	107	19	17	13	19
1600	91	73	50	40	54
2048	151	94	84	66	98

Table 10: Number of empty alignments across all the $5,103$ sentences in our test sets for NLLB (top) and TOWERBASE (bottom) models.

		ds-BLEU		COMET	
		FT-U	FT-G	FT-U	FT-G
NLLB	sent	-0.8 (0.00)	-0.8 (0.01)	-	-0.3 (0.01)
	256	7.6 (0.00)	8.0 (0.00)	13.0 (0.00)	13.8 (0.00)
	512	26.3 (0.00)	27.0 (0.00)	33.4 (0.00)	33.8 (0.00)
	768	33.5 (0.00)	28.0 (0.00)	39.0 (0.00)	27.8 (0.00)
	1024	33.5 (0.00)	21.2 (0.00)	41.9 (0.00)	22.6 (0.00)
TOWERBASE	sent	2.4 (0.00)	2.6 (0.00)	0.7 (0.00)	0.7 (0.00)
	256	2.1 (0.00)	1.6 (0.00)	2.3 (0.00)	2.3 (0.00)
	512	2.1 (0.00)	2.0 (0.01)	2.4 (0.00)	2.6 (0.00)
	768	3.5 (0.00)	3.2 (0.00)	3.7 (0.00)	3.7 (0.00)
	1024	5.6 (0.00)	5.0 (0.00)	5.6 (0.00)	5.5 (0.00)
	1200	5.4 (0.00)	5.1 (0.00)	5.5 (0.00)	5.4 (0.00)
	1600	8.4 (0.00)	8.3 (0.00)	10.1 (0.00)	10.3 (0.00)
	2048	6.8 (0.00)	6.1 (0.00)	8.8 (0.00)	7.9 (0.00)

Table 11: Average difference (and p-values) in ds-BLEU or $100\times$COMET between fine-tuned models (FT) and the corresponding pretrained models NLLB (top) and TOWERBASE (bottom). U and G denote the corpora **TED-U** and **TED-G** respectively. - for p-values > 0.05. Positive values indicate that the fine-tuned model improves over the baseline.

		2014	2015	2016	2017
sent	FT	43.8 (0.98)	44.0 (0.99)	40.4 (1.00)	41.3 (1.00)
	Unif	42.4 (0.98)	42.5 (0.99)	39.7 (1.00)	40.2 (1.00)
	SHAPE	40.7 (0.97)	41.5 (0.98)	38.3 (1.00)	39.4 (1.00)
256	FT	43.5 (0.98)	43.1 (0.99)	40.4 (1.00)	40.8 (1.00)
	Unif	43.9 (0.98)	43.2 (0.99)	39.7 (1.00)	40.6 (1.00)
	SHAPE	43.3 (0.98)	43.2 (0.99)	39.8 (1.00)	40.2 (0.99)
512	FT	43.5 (0.98)	43.4 (0.98)	40.2 (1.00)	40.4 (1.00)
	Unif	43.8 (0.98)	43.8 (0.99)	39.8 (1.00)	41.0 (1.00)
	SHAPE	40.6 (0.91)	40.5 (0.92)	37.0 (0.9)	40.4 (0.98)
768	FT	36.6 (0.87)	36.4 (0.88)	35.3 (0.92)	35.6 (0.93)
	Unif	41.8 (0.95)	39.1 (0.88)	38.1 (0.95)	39.4 (0.97)
	SHAPE	34.0 (0.75)	34.9 (0.77)	33.8 (0.84)	34.9 (0.85)
1024	FT	28.6 (0.70)	29.1 (0.75)	28.9 (0.80)	28.7 (0.79)
	Unif	36.1 (0.81)	38.7 (0.87)	34.9 (0.87)	37.2 (0.92)
	SHAPE	32.4 (0.73)	30.9 (0.69)	30.8 (0.75)	28.0 (0.69)
1200	FT	25.2 (0.64)	25.8 (0.74)	24.1 (0.71)	24.3 (0.73)
	Unif	34.6 (0.80)	36.4 (0.82)	30.0 (0.74)	32.4 (0.80)
	SHAPE	27.2 (0.61)	30.3 (0.68)	24.6 (0.64)	23.9 (0.60)
1600	FT	18.2 (0.53)	19.3 (0.62)	19.2 (0.62)	19.6 (0.59)
	Unif	25.5 (0.59)	30.1 (0.70)	26.9 (0.68)	29.4 (0.72)
	SHAPE	22.0 (0.50)	21.7 (0.50)	22.1 (0.56)	20.6 (0.57)
2048	FT	15.4 (0.49)	12.4 (0.52)	16.7 (0.58)	14.8 (0.61)
	Unif	22.0 (0.51)	21.6 (0.50)	24.3 (0.60)	20.6 (0.53)
	SHAPE	18.7 (0.43)	15.1 (0.44)	14.6 (0.38)	13.4 (0.48)

		2014	2015	2016	2017
sent	FT	84	85	85	84
	Unif	82	84	84	83
	SHAPE	82	83	83	83
256	FT	81	82	82	81
	Unif	81	82	82	81
	SHAPE	80	82	81	80
512	FT	81	82	81	80
	Unif	81	82	81	81
	SHAPE	77	78	76	77
768	FT	69	69	70	69
	Unif	74	75	74	73
	SHAPE	68	69	68	68
1024	FT	58	59	60	58
	Unif	67	65	67	67
	SHAPE	60	62	61	59
1200	FT	56	56	55	53
	Unif	61	65	62	60
	SHAPE	55	59	55	54
1600	FT	50	49	49	49
	Unif	55	58	54	56
	SHAPE	51	52	50	48
2048	FT	47	42	48	45
	Unif	51	49	51	48
	SHAPE	46	43	47	44

Table 12: ds-BLEU (and brevity penalty) (left) and $100\times$COMET (right) scores for FT-NLLB-G (FT), UNIF-NLLB-G (Unif), and SHAPE-NLLB-U (SHAPE) trained on **TED-G** with target max source document length $M = 2048$.

		2014	2015	2016	2017
sent	FT	44.2 (0.99)	43.5 (0.99)	40.4 (1.00)	41.4 (1.00)
	Unif	40.1 (0.95)	40.4 (0.97)	38.0 (1.00)	38.1 (0.99)
	SHAPE	39.4 (0.97)	40.0 (0.98)	36.3 (1.00)	37.8 (0.99)
256	FT	43.2 (0.98)	42.8 (0.99)	39.7 (1.00)	40.5 (1.00)
	Unif	42.8 (0.98)	42.4 (0.99)	39.5 (1.00)	40.3 (1.00)
	SHAPE	42.0 (0.97)	42.5 (0.99)	39.6 (1.00)	39.4 (1.00)
512	FT	42.9 (0.98)	43.1 (0.99)	39.2 (1.00)	39.4 (1.00)
	Unif	43.4 (0.98)	43.0 (0.99)	39.8 (1.00)	40.5 (1.00)
	SHAPE	39.7 (0.89)	41.1 (0.94)	38.2 (0.95)	39.0 (0.98)
768	FT	43.5 (0.98)	43.0 (0.99)	39.4 (1.00)	39.8 (1.00)
	Unif	44.0 (0.98)	43.7 (0.99)	40.4 (1.00)	40.9 (1.00)
	SHAPE	39.6 (0.88)	41.4 (0.93)	37.4 (0.93)	36.8 (0.91)
1024	FT	42.6 (0.96)	42.6 (0.97)	39.2 (1.00)	39.6 (1.00)
	Unif	42.6 (0.96)	44.1 (0.98)	39.6 (1.00)	40.6 (0.99)
	SHAPE	40.3 (0.91)	42.8 (0.96)	36.8 (0.92)	37.8 (0.94)
1200	FT	38.3 (0.88)	39.3 (0.91)	36.3 (0.92)	36.4 (0.92)
	Unif	39.5 (0.89)	43.0 (0.97)	38.0 (0.95)	39.2 (0.98)
	SHAPE	36.9 (0.84)	37.2 (0.87)	32.6 (0.82)	34.3 (0.86)
1600	FT	31.5 (0.77)	30.4 (0.73)	30.9 (0.83)	30.3 (0.80)
	Unif	31.5 (0.72)	34.4 (0.82)	34.2 (0.88)	33.8 (0.84)
	SHAPE	28.0 (0.68)	29.4 (0.68)	31.1 (0.79)	31.7 (0.82)
2048	FT	27.4 (0.69)	24.0 (0.63)	26.7 (0.75)	23.6 (0.71)
	Unif	30.2 (0.68)	25.3 (0.60)	31.5 (0.79)	28.0 (0.68)
	SHAPE	26.1 (0.63)	27.9 (0.66)	26.8 (0.69)	26.9 (0.74)

		2014	2015	2016	2017
sent	FT	84	85	85	84
	Unif	82	83	83	83
	SHAPE	81	82	82	82
256	FT	80	81	81	81
	Unif	80	81	81	81
	SHAPE	79	81	81	80
512	FT	80	81	81	81
	Unif	81	82	81	81
	SHAPE	75	79	79	78
768	FT	80	81	81	81
	Unif	81	82	82	81
	SHAPE	75	78	75	75
1024	FT	78	79	78	78
	Unif	80	81	81	80
	SHAPE	74	78	75	73
1200	FT	71	73	70	69
	Unif	76	79	76	75
	SHAPE	68	70	69	66
1600	FT	61	60	61	60
	Unif	62	61	61	64
	SHAPE	57	59	62	59
2048	FT	57	53	57	54
	Unif	59	57	57	54
	SHAPE	51	52	57	53

Table 13: ds-BLEU (and brevity penalty) (left) and 100×COMET (right) scores for FT-NLLB-U (FT) UNIF-NLLB-U (Unif), and SHAPE-NLLB-U (SHAPE) trained on **TED-U** with target max source document length $M = 2048$.

		2014	2015	2016	2017
sent	FT	46.5 (0.98)	45.1 (0.99)	42.3 (1.00)	41.0 (1.00)
	Unif	46.5 (0.98)	45.0 (0.99)	42.3 (1.00)	41.1 (1.00)
	SHAPE	46.4 (0.98)	45.2 (0.99)	42.4 (1.00)	41.2 (1.00)
256	FT	44.6 (0.98)	45.1 (0.99)	42.3 (1.00)	41.9 (1.00)
	Unif	44.5 (0.99)	45.1 (0.99)	42.2 (1.00)	41.8 (1.00)
	SHAPE	46.2 (0.98)	45.2 (0.99)	42.4 (1.00)	42.1 (1.00)
512	FT	43.7 (0.98)	45.0 (1.00)	41.4 (1.00)	41.6 (1.00)
	Unif	43.7 (0.98)	44.8 (1.00)	41.4 (1.00)	41.4 (1.00)
	SHAPE	45.5 (0.98)	45.1 (0.99)	41.6 (1.00)	41.6 (1.00)
768	FT	44.0 (0.98)	44.2 (0.99)	40.3 (1.00)	40.7 (1.00)
	Unif	44.0 (0.98)	44.2 (0.99)	40.2 (1.00)	40.7 (1.00)
	SHAPE	45.6 (0.98)	44.4 (0.99)	41.3 (1.00)	41.5 (1.00)
1024	FT	42.7 (0.98)	40.6 (0.99)	38.9 (1.00)	40.4 (1.00)
	Unif	42.2 (0.97)	42.6 (0.99)	39.4 (1.00)	40.1 (1.00)
	SHAPE	44.5 (0.98)	40.9 (0.99)	39.2 (1.00)	39.7 (1.00)
1200	FT	42.6 (0.98)	43.0 (0.99)	38.9 (1.00)	40.8 (1.00)
	Unif	42.5 (0.98)	42.7 (0.99)	39.3 (1.00)	40.6 (1.00)
	SHAPE	44.0 (0.98)	42.8 (0.99)	39.3 (1.00)	40.6 (1.00)
1600	FT	42.0 (0.97)	41.1 (0.98)	37.0 (1.00)	38.7 (1.00)
	Unif	42.0 (0.97)	40.0 (0.98)	37.5 (1.00)	37.3 (1.00)
	SHAPE	42.2 (0.97)	40.6 (0.98)	38.5 (1.00)	39.2 (0.99)
2048	FT	33.0 (0.99)	32.5 (0.99)	31.6 (1.00)	29.2 (0.97)
	Unif	33.1 (0.99)	33.7 (0.99)	32.2 (0.98)	26.5 (0.97)
	SHAPE	34.1 (0.97)	35.1 (0.98)	32.1 (1.00)	30.2 (0.97)

		2014	2015	2016	2017
sent	FT	85	86	85	85
	Unif	85	86	85	85
	SHAPE	85	86	85	85
256	FT	83	84	83	83
	Unif	83	84	83	83
	SHAPE	83	84	83	83
512	FT	82	84	83	82
	Unif	82	84	83	82
	SHAPE	82	84	83	82
768	FT	82	83	83	82
	Unif	82	83	83	82
	SHAPE	82	84	83	82
1024	FT	81	81	82	81
	Unif	81	82	83	81
	SHAPE	81	81	83	81
1200	FT	78	82	81	81
	Unif	78	82	81	81
	SHAPE	80	82	82	81
1600	FT	80	79	79	80
	Unif	79	78	79	78
	SHAPE	79	79	80	79
2048	FT	68	69	68	64
	Unif	69	70	70	62
	SHAPE	65	73	70	68

Table 14: ds-BLEU (and brevity penalty) (left) and 100×COMET (right) scores for FT-TOWER-G (FT), UNIF-TOWER-G (Unif) and SHAPE-TOWER-G (SHAPE) trained on **TED-G** with target max source document length 2048 ($M = 4096$).

		2014	2015	2016	2017
sent	FT	46.2 (0.99)	45.1 (0.99)	42.1 (1.00)	40.8 (1.00)
	Unif	46.4 (0.98)	45.1 (0.99)	42.2 (1.00)	40.9 (1.00)
	SHAPE	46.3 (0.98)	45.2 (0.99)	42.4 (1.00)	41.0 (1.00)
256	FT	46.3 (0.98)	45.1 (0.99)	42.3 (1.00)	41.8 (1.00)
	Unif	44.5 (0.98)	45.2 (0.99)	42.3 (1.00)	41.9 (1.00)
	SHAPE	46.0 (0.98)	45.3 (0.99)	42.4 (1.00)	42.0 (1.00)
512	FT	44.4 (0.98)	44.9 (0.99)	41.2 (1.00)	41.6 (1.00)
	Unif	43.0 (0.98)	45.0 (0.99)	41.4 (1.00)	41.6 (1.00)
	SHAPE	44.5 (0.98)	45.0 (0.99)	41.3 (1.00)	41.6 (1.00)
768	FT	44.1 (0.98)	44.6 (0.99)	41.1 (1.00)	40.8 (1.00)
	Unif	43.2 (0.99)	44.5 (0.99)	41.4 (1.00)	41.0 (1.00)
	SHAPE	44.2 (0.99)	44.5 (0.99)	41.2 (1.00)	41.7 (1.00)
1024	FT	43.9 (0.97)	41.6 (0.99)	39.6 (1.00)	39.7 (1.00)
	Unif	42.7 (0.98)	43.8 (0.99)	39.7 (1.00)	39.6 (1.00)
	SHAPE	44.1 (0.98)	42.9 (0.99)	39.8 (1.00)	39.5 (1.00)
1200	FT	43.2 (0.97)	42.9 (0.99)	39.5 (1.00)	40.8 (1.00)
	Unif	43.2 (0.98)	43.1 (0.99)	38.2 (1.00)	40.9 (1.00)
	SHAPE	43.8 (0.98)	43.2 (0.99)	39.2 (1.00)	39.7 (1.00)
1600	FT	41.6 (0.95)	40.5 (0.97)	37.6 (1.00)	39.8 (0.99)
	Unif	41.1 (0.97)	41.6 (0.98)	36.5 (1.00)	38.0 (1.00)
	SHAPE	42.6 (0.96)	42.1 (0.98)	37.3 (1.00)	40.5 (1.00)
2048	FT	34.4 (0.96)	35.3 (0.97)	31.2 (0.99)	28.2 (0.96)
	Unif	34.6 (0.97)	35.5 (0.98)	30.2 (1.00)	30.9 (0.97)
	SHAPE	35.2 (0.97)	34.6 (0.95)	31.9 (0.99)	31.5 (0.96)

		2014	2015	2016	2017
sent	FT	85	86	85	85
	Unif	85	86	85	85
	SHAPE	85	86	85	86
256	FT	83	84	83	83
	Unif	83	84	83	82
	SHAPE	83	84	83	82
512	FT	82	84	83	82
	Unif	82	84	83	82
	SHAPE	82	84	83	82
768	FT	82	83	83	82
	Unif	82	84	83	82
	SHAPE	82	83	83	82
1024	FT	81	81	83	81
	Unif	81	83	83	81
	SHAPE	81	82	83	81
1200	FT	80	82	81	81
	Unif	80	82	81	81
	SHAPE	80	82	81	80
1600	FT	79	79	80	80
	Unif	80	80	78	79
	SHAPE	79	81	78	80
2048	FT	68	72	69	64
	Unif	70	72	69	67
	SHAPE	70	73	70	68

Table 15: ds-BLEU (and brevity penalty) (left) and 100×COMET (right) for FT-TOWER-U (FT), UNIF-TOWER-U (Unif) and SHAPE-TOWER-U (SHAPE) trained on **TED-U** with target max source document length 2048 ($M = 4096$).

	TED-U		**TED-G**		FT	Unif	SHAPE
	FT vs Unif	FT vs SHAPE	FT vs Unif	FT vs SHAPE	U vs G	U vs G	U vs G
sent	-0.1 (0.20)	**-0.2 (0.01)**	-0.0 (0.60)	-0.1 (0.18)	**-0.1 (0.05)**	-0.1 (0.19)	-0.1 (0.36)
256	0.5 (0.32)	-0.0 (0.87)	0.1 (0.22)	-0.5 (0.24)	0.5 (0.32)	0.1 (0.21)	-0.1 (0.48)
512	0.3 (0.55)	-0.1 (0.10)	0.1 (0.08)	-0.6 (0.23)	0.1 (0.87)	-0.1 (0.82)	-0.4 (0.28)
768	0.2 (0.46)	-0.2 (0.67)	0.0 (0.84)	**-0.9 (0.05)**	0.3 (0.11)	0.2 (0.65)	-0.4 (0.19)
1024	-0.2 (0.82)	-0.3 (0.49)	-0.4 (0.44)	-0.5 (0.35)	0.6 (0.29)	0.4 (0.34)	0.5 (0.38)
1200	0.2 (0.44)	0.1 (0.82)	0.0 (0.70)	-0.4 (0.06)	0.3 (0.24)	0.1 (0.71)	-0.2 (0.50)
1600	0.6 (0.45)	**-0.7 (0.01)**	0.4 (0.58)	-0.5 (0.40)	0.1 (0.84)	0.0 (1.00)	0.4 (0.49)
2048	-0.5 (0.61)	-1.0 (0.22)	0.2 (0.84)	-1.3 (0.19)	0.7 (0.46)	1.3 (0.13)	0.4 (0.44)
sent	-0.0 (0.52)	-0.0 (0.93)	0.0 (0.38)	0.0 (0.60)	-0.0 (0.95)	0.0 (0.24)	0.0 (0.56)
256	0.0 (0.60)	0.0 (0.85)	0.0 (0.73)	-0.1 (0.07)	-0.0 (0.63)	-0.1 (0.46)	**-0.2 (0.05)**
512	-0.1 (0.15)	**-0.2 (0.04)**	0.1 (0.16)	-0.1 (0.27)	-0.2 (0.09)	0.0 (0.76)	-0.0 (0.62)
768	-0.1 (0.29)	0.1 (0.28)	-0.1 (0.34)	**-0.3 (0.00)**	0.0 (0.87)	0.0 (0.69)	**-0.4 (0.00)**
1024	-0.6 (0.22)	-0.5 (0.32)	-0.3 (0.46)	-0.2 (0.09)	0.1 (0.45)	0.3 (0.10)	0.4 (0.42)
1200	0.1 (0.69)	0.2 (0.51)	0.1 (0.50)	-0.4 (0.16)	0.1 (0.61)	0.0 (0.93)	-0.4 (0.12)
1600	0.2 (0.69)	-0.4 (0.27)	0.6 (0.37)	0.1 (0.70)	-0.2 (0.62)	0.2 (0.73)	0.4 (0.49)
2048	-0.9 (0.40)	-1.3 (0.16)	-0.2 (0.84)	-1.7 (0.13)	0.9 (0.50)	1.5 (0.10)	0.4 (0.47)

Table 16: Average difference (and p-values) in **ds-BLEU** (top) evaluated on full TED talks and 100×**COMET** (bottom) evaluated on realigned sentences for TOWERBASE. Left and middle: paired comparison between the original fine-tuning (FT), UNIFPE and SHAPE on **TED-U** and **TED-G** respectively. Right: differences between fine-tuning on TED-U (U) and TED-G (G). A positive value indicates that in the comparison pair, the translation of the first item achieves higher scores than that of the second. Significant differences with p-values < 0.05 are in bold.

Investigating the translation capabilities of Large Language Models trained on parallel data only

Javier García Gilabert, Carlos Escolano, Aleix Sant,
Francesca De Luca Fornaciari, Audrey Mash, Xixian Liao, Maite Melero
Barcelona Super Computing Center (BSC)
{javier.garcia1,carlos.escolano,aleix.santsavall,
francesca.delucafornaciari,audrey.mash,xixian.liao,maite.melero}@bsc.es

Abstract

In recent years, Large Language Models (LLMs) have demonstrated exceptional proficiency across a broad spectrum of Natural Language Processing (NLP) tasks, including Machine Translation. However, previous methods predominantly relied on iterative processes such as instruction fine-tuning or continual pre-training, leaving unexplored the challenges of training LLMs solely on parallel data. In this work, we introduce PLUME (**P**arallel **L**ang**u**age **M**od**e**l), a collection of three 2B LLMs[1] featuring varying vocabulary sizes (32k, 128k, and 256k) trained exclusively on Catalan-centric parallel examples. These models perform comparably to previous encoder-decoder architectures on 16 supervised translation directions and 56 zero-shot ones. Utilizing this set of models, we conduct a thorough investigation into the translation capabilities of LLMs, probing their performance, the role of vocabulary size, the impact of the different elements of the prompt, and their cross-lingual representation space. We find that larger vocabulary sizes improve zero-shot performance and that different layers specialize in distinct aspects of the prompt, such as language-specific tags. We further show that as the vocabulary size grows, a larger number of attention heads can be pruned with minimal loss in translation quality, achieving a reduction of over 64.7% in attention heads.

We release our code at
https://github.com/
projecte-aina/Plume

1 Introduction

Neural Machine Translation (NMT) has traditionally relied on encoder-decoder architectures, where an encoder processes the source sentence and a decoder generates the target sentence based on the encoder's output. However, recent advancements have moved away from this paradigm, with the introduction of decoder-only Large Language Models (LLMs). In these models, the source sentence acts as a prompt, eliminating the need for a conventional encoder.

With the rise of LLMs, research has increasingly focused on adapting these models for translation tasks by using techniques such as prompt-tuning (Zhang et al., 2023), instruction-finetuning (Xu et al., 2024), or continual pretraining (Rei et al., 2022a). While these methods have shown impressive results, they open new questions about the performance of LLMs when trained exclusively on parallel data, and therefore, the possibility of having models that are trained directly on the task of machine translation. Additionally, the majority of these models are trained predominantly on English-centric-corpora.

To address these questions, our paper proposes a new approach consisting of training LLMs solely on parallel corpora to evaluate their efficacy in machine translation (MT). Our investigation revolves around questions such as: How does an LLM trained exclusively on parallel data perform? And how does the model leverage prompt information to ensure accurate translations?

Our contributions are twofold: Firstly, we introduce PLUME (**P**arallel **L**ang**u**age **M**od**e**l), an innovative ensemble comprising three multilingual 2B LLMs, trained from scratch on Catalan-centric parallel data. Each model has a different vocabulary size (32k, 128k and 256k). All models are proficient in 16 supervised translation directions, as well as 56 zero-shot translation directions. Results show comparable results to previous encoder-decoder architectures of similar size.

Secondly, to understand how these models work, we study how they utilize contextual information

[1]We release our models on HuggingFace: Plume 32k, Plume 128k and Plume 256k.

Proceedings of Machine Translation Summit XX Volume 1, pages 24–53, June 23-27, 2025

across different layers to execute translation tasks effectively. Our experiments show distinctive attention patterns associated with the different parts of the prompt, and how they vary through the different attention blocks. We also observe how languages use the source tag information differently, leading to a large performance variability when this token is missing. As a byproduct, we propose a strategy to remove attention heads with minimal performance loss and study how vocabulary size impacts the appearance of redundant heads. Finally, we study the cross-lingual space learned by the models and how it progresses through the model's attention blocks.

2 Related work

Neural Machine Translation (NMT) has predominantly relied on encoder-decoder architectures (Cho et al., 2014; Bahdanau et al., 2015; Sutskever et al., 2014). These methods have proven effective by conditioning language models to generate translations that accurately retain the meaning of the source sentence. Moreover, these systems are easily extendable to multilingual scenarios, enabling zero-shot translation between language pairs that have not been seen together during training (Firat et al., 2016; Wu et al., 2016).

Over the years, some approaches to NMT have dropped the traditional encoder-decoder setup to adopt decoder-only architectures (Fonollosa et al., 2019; He et al., 2018). Although these methods showed promise, they did not become the standard due to issues with context loss and hallucinations (Fu et al., 2023).

Recent advancements in training Large Language Models (LLMs) (Touvron et al., 2023; Jiang et al., 2023; Gemma Team et al., 2024; Abdin et al., 2024), including techniques like scaling and Rotary Embeddings (Su et al., 2024b), have significantly enhanced the ability of decoder-only architectures to handle long contexts of hundreds or even thousands of tokens. Consequently, several studies have proposed leveraging pretrained LLMs for NMT through continual pretraining and instruction tuning (Alves et al., 2024; Xu et al., 2024; Yang et al., 2023). These methods have demonstrated results comparable to traditional encoder-decoder systems, while also supporting multiple translation directions.

However, training and adapting these systems to various languages remains challenging (Ali et al., 2024). Creating a vocabulary that accurately represents all supported languages can lead to performance disparities of up to 68% on some downstream tasks. Additionally, interpretability methods have gained popularity in order to understand better how models utilize provided information and to guide further improvements (Voita et al., 2019b,a; Ferrando et al., 2024).

3 Methodology

3.1 Catalan-Centric Dataset

In order to study zero-shot translation using a decoder-only architecture, we employ a Catalan-centric dataset. This dataset pairs Catalan sentences with their counterparts in one of eight other languages: Spanish, French, Italian, Portuguese, Galician, German, English, and Basque. Specifically, for each language, we include translation directions both to Catalan (xx→ca) and from Catalan (ca→xx). The dataset consists of 783.6M sentences and 30.9 billion words. We show in Table 1 the number of sentences and number of words per language pair in the created dataset.

Pair	N sentences	N words
ca SYN ↔ de	187,483,456	6,847,140,698
ca ↔ de	12,516,544	603,121,312
ca SYN ↔ it	181,034,146	6,526,304,128
ca ↔ it	18,965,862	577,243,404
ca ↔ es	171,907,026	8,252,262,032
ca SYN ↔ pt	62,858,532	2,429,548,286
ca ↔ pt	12,319,262	504,959,082
ca ↔ en	60,046,068	2,429,961,320
ca ↔ fr	37,269,716	1,114,635,790
ca SYN ↔ eu	17,998,782	749,042,034
ca ↔ eu	2,091,356	61,237,122
ca SYN ↔ gl	11,434,180	531,773,730
ca ↔ gl	7,713,022	263,280,596
Total	783,637,952	30,890,509,534

Table 1: Number of sentences and words for each language pair. We label languages with their BCP-47 language code. SYN means synthetic data generated on the source side for the ca-xx direction.

Data preprocessing All data is first filtered using LaBSE (Feng et al., 2022) to embed both source and target sentences then compute a cosine similar-

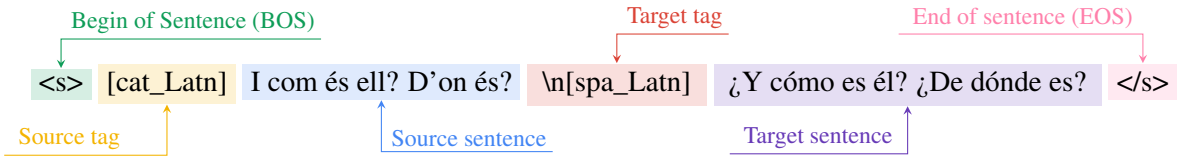


Figure 1: Prompt strategy used to train PLUME.

ity score between both[2]. Off-target translations are removed using the *Lingua*[3] library[4]. Following the filtering process, the data undergoes deduplication and punctuation normalization using the *Bifixer* library (Ramírez-Sánchez et al., 2020). Further details about the dataset are provided in Appendix A.

3.2 Tokenizer

Prior studies have shown that vocabulary overlap plays a crucial role in zero-shot translation for encoder-decoder architectures (Stap et al., 2023; Tan and Monz, 2023). More related to our work concerning tokenizer size in decoder-only architectures is the study by Ali et al. (2024), who found that larger vocabulary sizes lead to improved downstream performance in multilingual settings. The main difference is that our focus is in Multilingual Neural Machine Translation (MNMT) while Ali et al. (2024) focused on more general multilingual tasks (Natural language inference, Question Answering, etc.).

To investigate the impact of vocabulary sharing on zero-shot MNMT for decoder-only architectures, we train 3 tokenizers using BPE (Sennrich et al., 2016) from the *Huggingface tokenizer* library (Moi and Patry, 2023) with different vocabulary sizes; 32k, 128k, and 256k. Regarding the training data used to train the tokenizer, recent work has shown that while NMT performance is relatively robust to language imbalance, better performance is often achieved when languages are more equally represented in the training data (Zhang et al., 2022). In this work, we equally sample Romance languages and we oversample English, Basque, and German to avoid underrepresenting these languages and to achieve near parity (Petrov et al., 2024) and fertility among all language pairs. Average fertility (average of fertility per each language) per vocabulary size as well as the number of tokens in the dataset are shown in Table 2 [5]. More details about tokenizer experiments can be found in Appendix B.

	Avg. Fertility	N tokens
PLUME 32k	1.77	54.7B
PLUME 128k	1.52	46.8B
PLUME 256k	1.44	44.6B

Table 2: Fertility and number of tokens in the dataset grouped by vocabulary size.

3.3 Model

We trained one model for each of our three tokenizers using the same architecture as GEMMA 2B[6] (Gemma Team et al., 2024) to train a 2 billion parameter, transformer-based, decoder-only model. Following the scaling law proposed by (Hoffmann et al., 2022), each model was trained on 30.9 billion words, corresponding to 54.7, 46.8, and 44.6 billion tokens for vocabularies of 32k, 128k, and 256k respectively. Details about the model size and model architecture are shown in Table 3.

Hyper-Parameter	Value
Hidden size	2048
Layers	18
Feedforward size	16384
Attention-Heads	8
Head size	256
Num KV Heads	1
Max Seq Length	2048
Position Embeddings	Rotary (Su et al., 2024a)
Rope Theta	10000
Precision	float-32
RMSNorm ϵ	1e-06
Activation	GeGLU (Shazeer, 2020)

Table 3: Model architecture of PLUME models.

[2]We use a cosine similarity threshold of 0.75 for LaBSE filtering.

[3]https://github.com/pemistahl/lingua-py

[4]We use a threshold of 0.5 for the language probability score.

[5]We compute the number of tokens as Average Fertility * Number of words in the dataset. The number of words is 30,890,509,534.

[6]https://huggingface.co/google/gemma-2b

3.4 Training

We train all PLUME models with a context window of 2048 tokens, utilizing the Adam optimizer (Kingma and Ba, 2015) and the causal language modeling objective. The learning rate is warmed up from 1×10^{-7} to a maximum of 3×10^{-4} over the first 2000 steps. We apply a weight decay of 0.1 and a gradient clipping of 1.0. During training, we set an effective batch size of 81,920 tokens per gradient step distributed over 40 NVIDIA H100-64GB GPUs using the Deepspeed framework[7].

Note that the main focus of this study is to understand how LLMs perform translation. Thus, PLUME models are not trained for state-of-the-art performance on MNMT. A more detailed description of the training configuration can be found in Appendix C.

Formatting Figure 1 presents an example of a formatted sentence for the Catalan to Spanish translation direction. During batching, we concatenate formatted sentences up to a context length of 2048 tokens, mixing different translation directions within a single batch. Padding is added to fill out the remainder of the sequence.

3.5 Evaluation

To compute reference-based translation quality we use COMET-22 (Rei et al., 2022a) and BLEU (Papineni et al., 2002) metrics on the FLORES-200 devtest (NLLB Team et al., 2022) and NTREX-101 (Federmann et al., 2022) datasets. We additionally report CHRF (Popović, 2015) and COMET-KIWI-22 (Rei et al., 2022b) in appendix G. We use TOWEREVAL[8] (Alves et al., 2024) to compute all the evaluation metrics. For inference, we use beam search decoding with a beam size of 5 and limiting the translation length to 512 tokens.

We compare PLUME models with the following bilingual and multilingual models.

- NLLB (NLLB Team et al., 2022): A transformer encoder-decoder model that supports 202 languages. We use the 600 million, the 1.3 billion, and the 3.3 billion parameter variants.
- Bilingual models BSC: Transformer encoder-decoder models, trained from scratch on language pairs that include Catalan. These models were developed as part of the Aina Project[9] and follow the Transformer-XLarge architecture (Subramanian et al., 2021) featuring 500 million parameters in total.

It is important to note that NLLB has seen parallel data for our zero-shot directions, therefore zero-shot only describes the condition in PLUME models. Our setup is designed to study the potential of a decoder-only architecture to perform zero-shot translation, specifically using Catalan as the pivot language.

4 Results

Table 4 shows results for all PLUME models aggregated by supervised and zero-shot directions. The PLUME 32k, 128k and 256k variants perform equally well in supervised directions, achieving similar BLEU and COMET scores for both NTREX and FLORES-200 datasets. In supervised directions, PLUME models demonstrate competitive performance, matching the COMET scores of the Bilingual BSC models and achieving scores comparable to the NLLB variants.

In zero-shot directions, the PLUME models exhibit a decline in performance compared to supervised directions. However, the decline is more pronounced in the BLEU scores than in the COMET scores, indicating that the overall quality remains relatively robust. Specifically, the PLUME 256k variant achieves a COMET score of 0.84 on the FLORES-200 dataset and 0.81 on the NTREX dataset, which, although lower than its supervised performance, still demonstrates its zero-shot translation capabilities when training using only Catalan as the bridge language.

Larger vocabulary sizes improve zero-shot translation. The results in Table 4 show that higher vocabulary sizes consistently yield better zero-shot capabilities. Specifically, the PLUME 256k variant outperforms the 32k and 128k variants in zero-shot scenarios for both FLORES-200 and NTREX datasets.

To further understand the influence of the vocabulary size in zero-shot translation quality, we calculated the vocabulary overlap (Tan and Monz, 2023) for each zero-shot translation direction as follows:

$$Overlap = \frac{|V_{src} \cap V_{tgt}|}{|V_{tgt}|} \quad (1)$$

[7] https://www.deepspeed.ai/

[8] TOWEREVAL uses the sacreBLEU implementation to compute BLEU and CHRF metrics.

[9] https://huggingface.co/projecte-aina

	Supervised directions				Zero-shot directions			
	FLORES-200		NTREX		FLORES-200		NTREX	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
NLLB-3.3B	32.02	0.87	30.48	0.85	28.97	0.86	28.74	0.84
NLLB-1.3B	31.02	0.86	29.68	0.85	28.48	0.86	28.37	0.84
NLLB-600M	29.24	0.85	28.37	0.84	27.04	0.85	27.25	0.84
Bilinguals BSC	31.93	0.86	29.77	0.84	-	-	-	-
PLUME 32k	30.44	0.86	28.46	0.84	23.25	0.83	23.03	0.80
PLUME 128k	30.81	0.86	28.78	0.84	23.97	0.83	23.53	0.81
PLUME 256k	30.72	0.86	28.87	0.84	24.42	0.84	23.81	0.81

Table 4: Averaged BLEU and COMET scores on supervised and zero-shot directions for FLORES-200 devtest and NTREX.

where V_{src}, V_{tgt} are the set of unique words in the source and target language vocabulary respectively. We show the correlation between vocabulary overlap and both BLEU and COMET for zero-shot directions in table 5. On average there is a positive correlation between the vocabulary overlap and the translation quality of 0.3 for BLEU and 0.57 for COMET, which diminishes as vocabulary size increases. This suggests that vocabulary overlap between the source and target languages further helps explain zero-shot performance, particularly for smaller vocabulary sizes.

	PLUME 32k	PLUME 128k	PLUME 256k
BLEU	0.351	0.280	0.255
COMET	0.593	0.588	0.538

Table 5: Correlation between vocabulary overlap and BLEU, COMET metrics for different vocabulary sizes in zero-shot directions.

4.1 Understanding translation with an LLM

Our goal is to understand how an LLM performs translation. We start by examining which parts of the prompt the model focuses on. This helps us determine the most important attention heads for each section of the prompt. Then, we study the model's cross-lingual representation space by extracting contextualized token embeddings.

4.2 Attention

For each attention head, we assess its importance by calculating coverage as defined by (Tu et al., 2016). Originally, coverage was proposed for encoder-decoder attention and refers to the total attention a source token receives from target tokens. We adapt coverage for masked-self attention. Given a set of prompt's tokens I, the coverage formula for a single sentence is defined as:

$$\text{cov}_I(\text{head}) = \sum_{j \in J} \left(\sum_{i \in I} \alpha_{i,j} \right)^2 \quad (2)$$

where $\alpha_{i,j}$ denotes the attention weight from token i to token j and J represent the set of the decoded (target) tokens.

Each coverage metric is computed and averaged over the FLORES-200 devtest for each head in the model and for each translation direction. To under-

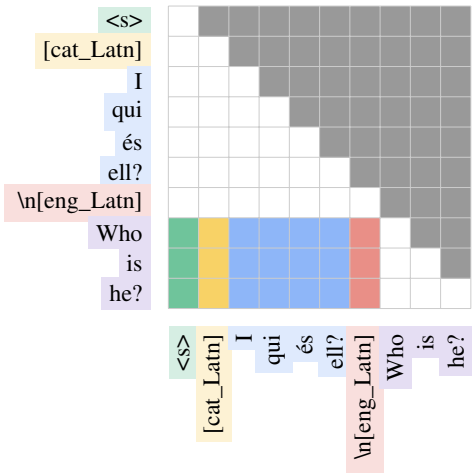


Figure 2: Illustration of the regions in the attention matrix used to compute coverage for each part of the prompt. We show the cross-attention regions between decoded tokens and the BOS, source tag, source sentence and target tag tokens in green, yellow, blue, and red, respectively.

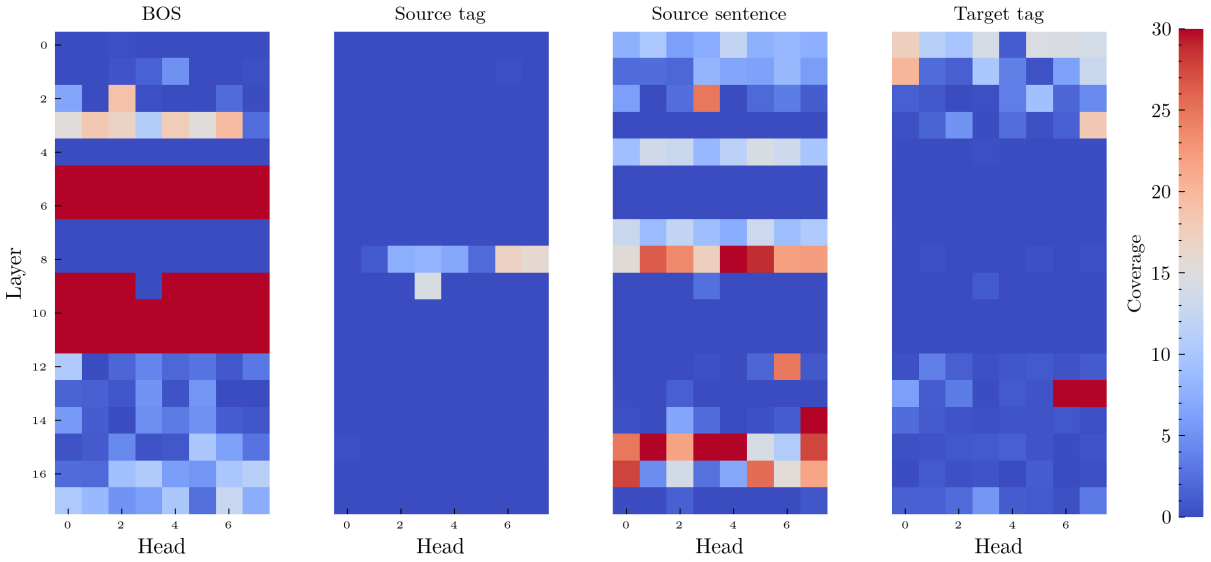


Figure 3: Coverage evaluating on FLORES-200 devtest using PLUME 32k. Each heatmap for each studied part of the prompt shows the coverage scores for each layer (on the vertical axis) and for each head (on the horizontal axis) in the model.

stand which part of the prompt the model is focusing on in each head we study coverage separately for different parts of the prompt: BOS, source tag, source sentence and target tag. Figure 2 shows a graphical illustration of the regions in the attention matrix that are used to compute coverage based on the part of the prompt.

In Figure 3, we show the average coverage across all translation directions for each part of the prompt, employing PLUME 32k. We note that heads within the same layer generally exhibit similar coverage patterns. Future work may investigate how these patterns arise and how they are related to the usage of Multi-Query attention[10] (Shazeer, 2019).

We find that source tag is the part of the prompt with least coverage. However, BOS, source sentence and target tag tokens exhibit varying degrees of coverage with some coverage spikes in specific layers and heads. Interestingly, layers 5, 6, 10 and 11 show coverage uniquely for the BOS token which suggests that all attention mass is given to the BOS token, leaving the residual stream unchanged. This patterns have recently been observed in autoregressive language models and are named attention sink mechanisms (Xiao et al., 2024; Ferrando and Voita, 2024; Ferrando et al., 2024; Cancedda, 2024). For instance, Cancedda (2024) demonstrates that in Llama 2, the feed-forward blocks embed crucial information into the residual stream of the BOS token, enabling the attention sink mechanism to happen in subsequent layers. We show in appendix D the coverage heatmaps for PLUME 128k and 256k.

Source tag importance As previously pointed out, the source tag receives less attention than the other parts of the prompt. Specifically, it has an average coverage of 0.56 which is 3.7 times less coverage than the target token or 18.5 times less coverage than the BOS token. This motivates our next experiments which consist of evaluating PLUME models without indicating the source language. Specifically, we replace the source tag with another BOS token to mantain the same learned positional encodings and evaluate the model's performance on FLORES-200 devtest using BLEU. Table 6 shows the relative BLEU change with respect to the original model aggregated by language pair. Results show varying impacts across different language pairs when the source tag is omitted. For languages like English, French and Basque, the drop in BLEU scores is particularly significant. However, for other translation directions like Spanish and Catalan, the decrease in BLEU scores is negligible. This suggests that the model is more reliant on the source tag to represent certain languages, particularly those which are less related to the bridge language or those that the model has seen less during training.

[10]When we use Multi-Query attention with *num_kv_heads* set to 1, the keys and values are shared across all heads from a specific layer and is only the query that differs which may hinder the specialization of the heads.

	PLUME 32k	PLUME 128k	PLUME 256k
ca→xx	-1.80	-0.54	-0.83
es→xx	-0.43	0.23	-0.33
pt→xx	-8.13	-6.01	-5.54
gl→xx	-6.52	-4.18	-4.92
it→xx	-6.57	-10.79	-5.03
fr→xx	-13.16	-19.90	-17.63
de→xx	-7.54	-2.73	-6.73
en→xx	-19.83	-25.52	-20.03
eu→xx	-16.73	-11.03	-13.23
Avg.	-8.97	-8.94	-8.25

Table 6: Relative BLEU change with respect to PLUME models after ignoring the source tag. We label languages according to their BCP-47 language code (see Table 9 from Appendix A).

Regarding the vocabulary size, the model with a 256k vocabulary shows the smallest average decrease in BLEU scores, suggesting that a larger vocabulary may improve the model's representation of the source language.

Redundant heads Previous work on MNMT has shown that coverage is a good indicator for pruning cross attention heads in encoder-decoder architectures and can be used to improve model's efficiency without sacrificing the model's performance (Kim et al., 2021). Following Kim et al. (2021), we use coverage to prune heads in a decoder-only architecture to study the amount of redundant heads that are introduced as vocabulary size grows.

Specifically, we mask all attention heads within a specific layer that fall below a predetermined coverage threshold. We compute coverage per layer for a specific direction as follows:

$$\text{COV}_l = \phi(\sum_{i=1}^{H} \sum_{j \in \text{Pr}} \text{cov}_j\left(\text{head}_{l,i}\right)) \quad (3)$$

Pr = {BOS, Source tag, Source sentence, Target tag}

where COV_l represents the coverage of layer l, H is the total number of attention heads in the model, and Pr is a set that contains sets of tokens for each part of the prompt. Finally, ϕ is a MinMax Scaler used to normalize the metric between 0 and 1.

We use FLORES-200 devtest to evaluate the impact of masking heads per layer based on the coverage criterion (Equation 3). Figure 4 (left) illustrates the evolution of BLEU scores as we mask heads in PLUME 32k for the Spanish to Catalan direction (supervised). The right axis indicates the number of heads that are masked. We find that up to 64 heads can be masked without degrading the model's performance using a threshold of 0.2, representing 47.05% of the model's total heads. In Figure 4 (right), we show the cumulative coverage for the different parts of the prompt. We observe that for a threshold of 0.2, the masked heads represent 9.05%, 2.61%, 36% and 58.9% total coverage for the BOS, source tag, source sentence and target tag tokens respectively. This indicates that the majority of the masked heads are paying attention to the target tag token and to a lesser extent to the source sentence tokens. This suggests that these heads are less critical for maintaining translation quality. Specifically, when masking these 64 heads we are only using heads from layers 5, 6, 8, 9, 10, 11, 15, and 16 which are the layers with higher coverage for the BOS, source tag and source sentence tokens (see Figure 3). Regarding the source tag, we find that even though it is the part of the prompt with the lowest coverage, it is still useful for maintaining the translation quality. This observation aligns with our previous findings from section 4.2.

	PLUME 32k	PLUME 128k	PLUME 256k
de→ca	64	64	88
de→en	32	72	88
de→pt	64	64	88
es→ca	64	104	88
es→en	64	72	88
es→pt	64	104	88
fr→ca	64	64	88
fr→en	24	72	88
fr→pt	64	0	88
gl→ca	64	104	88
gl→en	24	72	88
gl→pt	64	64	88
it→ca	64	80	88
it→en	64	72	88
it→pt	64	0	88
Avg.	56.53	67.2	88

Table 7: Number of masked heads across different language pairs and vocabulary sizes such that BLEU drop is less than 2 points.

In Table 7, we report the number of heads that we

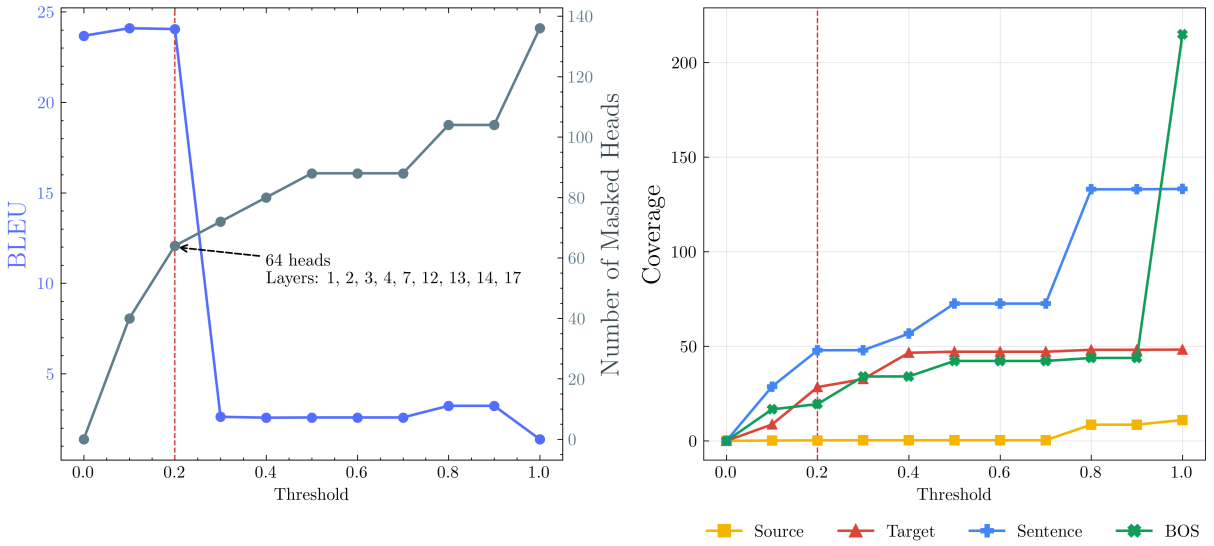


Figure 4: Impact of masking on BLEU score and number of masked heads across different coverage thresholds (left). Accumulated coverage of masked heads for source tag, target tag, source sentence, and BOS (right). Experiments are evaluated on the Spanish to Catalan direction.

can mask without losing more than 2 BLEU points for the translation directions from German (de), Spanish (es), French (fr), Galician (gl), and Italian (it) into Catalan (ca), English (en), and Portuguese (pt) for different vocabulary sizes. We find that for larger vocabulary sizes we can mask a higher number of heads without hurting the model's performance. Specifically, on average we can mask 41.56%, 49.41% and 64.7% of the model's heads for PLUME 32k, PLUME 128k and PLUME 256k respectively. Future work may investigate whether having more redundant heads is related with zero-shot translation, especially since larger vocabulary sizes appear to improve zero-shot translation capabilities.

4.3 Language subspaces

To further understand the multilingual capabilities of PLUME models, we study how different languages are represented within the model's internal representations by measuring the distances between language embeddings across layers and how do these representations differ across different vocabulary sizes.

Subspace distances We first extract sub-word tokens output by each layer in the Transformer. Specifically, we use the first 300 sentences from FLORES-200 devtest for each source language, denoted as s. These sentences are used to create translation prompts from s to each target language (300 * 8 = 2,400 prompts). For each prompt, we extract the token embeddings from each layer of the model and concatenate the consecutive tokens to form $\mathbf{H}_l^s$. Then, we apply singular value decomposition (SVD) on $\mathbf{H}_l^s$ after substracting the mean. We calculate pairwise distances among the 9 languages using the affine subspace for each language computed by the SVD, utilizing the Riemannian metric on the space of positive definite matrices described in (Chang et al., 2022), which is both symmetric and invariant to affine transformations.

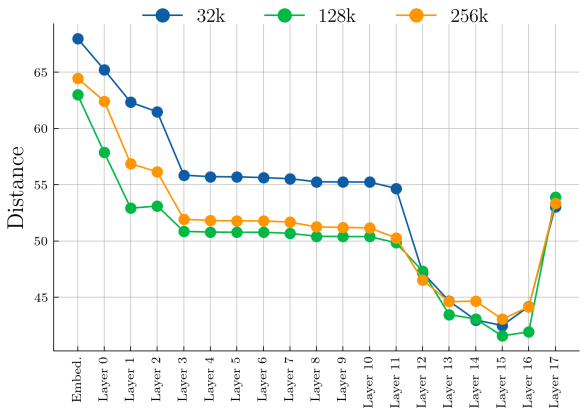


Figure 5: Mean distance between language subspaces grouped by vocabulary size. Additional plots grouped by languages and vocabulary sizes are included in Appendix E.

Figure 5 shows the mean distance between language subspaces in each layer. As we can see, the distance between language subspaces decreases with model depth. Initially, from the embeddings

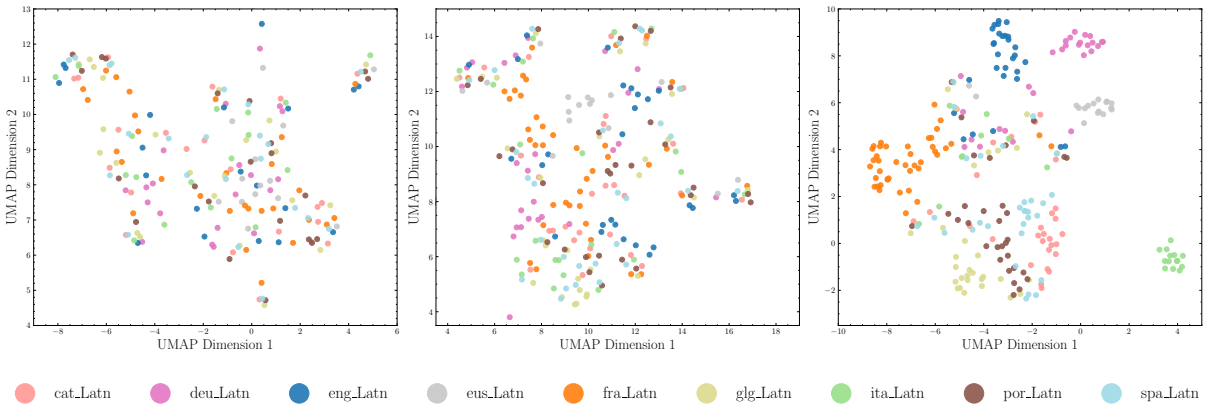


Figure 6: UMAP representations for token embeddings. From left to right: Representations at the embedding layer, the penultimate layer, and the last layer of PLUME 32k grouped by source language. See Appendix F for more additional plots.

layer to layer 0 we can observe a significant decrease of approximately 5.07%, and from layer 0 to layer 1, a further reduction of 7.23%. In middle layers (layers 3 to 11), distances are relatively stable and show minimal variations. This suggests that the model applies only minimal transformations to the representations along these layers. Interestingly, in layer 8 we can observe a small decrease in the distance of 0.05% which we hypothesize may be due to the model's attention focusing more on the source token at this layer (see Figure 3). As we move to the deeper layers, the distances continue to decrease, with a significant drop of approximately 8.88% from layer 11 to layer 12, a trend that persists through layers 12 to 16. However, in the last layer, there is a notable increase in distance by approximately 23.06%. These results align with previous work on encoder-only models, which suggest that in intermediate layers the model representations diverge more from the embedding layer representation and from the final layer. Both the embedding layer and the final layer are highly language-sensitive (Chang et al., 2022; Libovický et al., 2020; Pires et al., 2019).

Regarding the vocabulary size, as shown in Figure 5, we observe that for PLUME 32k the distance between embeddings are higher than PLUME 128k or PLUME 256k until layer 12, where distances become similar. This can be attributed to the higher vocabulary overlap between languages in PLUME 32k, where each embedding represents a more diverse concept, limiting its ability to learn language-agnostic representations which necessitates each embedding to represent more diverse concepts and suggests that a small vocabulary size might limit the model's ability to learn agnostic representations in earlier layers. In contrast, a larger vocabulary seems to help the model more readily disentangle language-specific features earlier in the network, allowing embeddings to remain closer.

Visualization In the previous subsection, we found that the distances between embeddings initially decrease, and the embedding space becomes narrower, then in the last layer, the embeddings spread out. To understand this phenomenon, we visualize the token embeddings using Uniform Manifold Approximation and Projection (UMAP)[11] (McInnes et al., 2018). We construct prompts from each source language to Galician. Token embeddings per layer are concatenated to form $\mathbf{P}_l^s$, then we apply UMAP to reduce the dimensionality of the representations.

Figure 6 shows the UMAP visualizations for token embeddings in the embedding layer and the two last layers of the model coloured by source language. As we can see, token embeddings remain language-neutral as they pass through the model until the last layer, where token embeddings group by source language. This suggests that the model must align embeddings cross-linguistically until reaching the last layer where it clusters by source language. This explains the distance of the last layer (see Figure 5). See Appendix F for additional plots[12] corresponding to each vocabulary size and each layer.

[11] We employ the cosine distance and we set the number of neighbours to 8 for computing UMAP's embeddings.

[12] Additionally, we include UMAP Spherical Voronoi diagrams as supplementary materials in the anonymous code: link (see Appendix F.1).

5 Conclusions

This work demonstrates the successful training of an LLM-based machine translation system from scratch using only parallel data. The achieved results are comparable to those of existing encoder-decoder architectures for supervised translation tasks. We identified that larger vocabulary sizes consistently improve translation quality across zero-shot directions, suggesting the potential benefits of experimenting with even larger or language-specific vocabularies.

Further analysis revealed that different layers focus on distinct aspects of the prompt, particularly the source language tag, which exhibits significant language variation. By employing an appropriate criterion, we achieved a performance reduction of less than 2 BLEU score while removing over 64.7% of attention heads. We also showed that with larger vocabularies, the model gains additional representational flexibility that allow for more heads to be pruned without significantly degrading performance.

Additionally, our exploration of the learned cross-lingual space demonstrates that languages get closer in the cross-lingual space as they get to deeper layers and highlight the layers with the most significant impact on the learned space.

This research opens doors for further investigation. We identified "sink heads" that primarily focus on the BOS token. Exploring their utility and relationship to the learned cross-lingual representations presents an opportunity for future work. Additionally, further research into the optimization of vocabulary size along model size could also lead to better NMT models.

6 Limitations

This study focused on understanding the capabilities of an LLM trained solely on parallel data, without aiming to achieve state-of-the-art translation quality or extensive language support. Here are some key limitations to consider when interpreting the results:

Data Scope: The experiment employed non-English centric data with a focus on Western, Latin-script languages. This approach aimed to isolate the impact of vocabulary size and overlap, but limits generalizability to languages with different scripts or historical connections. However, the inclusion of Basque, a non-Indo-European Subject-Object-Verb (SOV) language, provides valuable insights into the model's handling of structural variations.

Scalability: The study did not explore the impact of model scale and data availability on translation across diverse languages and scripts. Further research is necessary to understand how these factors influence performance in more complex settings.

These two main aspects will be considered as future work by studying the scalability of these architectures on both model size and translation directions.

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 DeepR3 (TED2021-130295B-C32) funded by MCIN/AEI/10.13039/501100011033 and European Union NextGeneration EU/PRTR.

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

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola,

Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Schulze Buschhoff, Charvi Jain, Alexander Arno Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2024. Tokenizer choice for llm training: Negligible or crucial? *Preprint*, arXiv:2310.08754.

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.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Nicola Cancedda. 2024. Spectral filters, dark signals, and attention sinks. *Preprint*, arXiv:2402.09221.

Tyler Chang, Zhuowen Tu, and Benjamin Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, pages 103–111. 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.

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.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models. *Preprint*, arXiv:2405.00208.

Javier Ferrando and Elena Voita. 2024. Information flow routes: Automatically interpreting language models at scale. *Preprint*, arXiv:2403.00824.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 866–875. The Association for Computational Linguistics.

José A. R. Fonollosa, Noe Casas, and Marta R. Costa-jussà. 2019. Joint source-target self attention with locality constraints. *Preprint*, arXiv:1905.06596.

Zihao Fu, Wai Lam, Qian Yu, Anthony Man-Cho So, Shengding Hu, Zhiyuan Liu, and Nigel Collier. 2023. Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder. *Preprint*, arXiv:2304.04052.

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. *Preprint*, arXiv:2403.08295.

Tianyu He, Xu Tan, Yingce Xia, Di He, Tao Qin, Zhibo Chen, and Tie-Yan Liu. 2018. Layer-wise coordination between encoder and decoder for neural machine translation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc.

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.

Zae Myung Kim, Laurent Besacier, Vassilina Nikoulina, and Didier Schwab. 2021. Do multilingual neural machine translation models contain language pair specific attention heads? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2832–2841, Online. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Anthony Moi and Nicolas Patry. 2023. Huggingface's tokenizers.

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.

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.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2024. Language model tokenizers introduce unfairness between languages. *Advances in Neural Information Processing Systems*, 36.

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.

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, 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.

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.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *Preprint*, arXiv:1911.02150.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

David Stap, Vlad Niculae, and Christof Monz. 2023. Viewing knowledge transfer in multilingual machine translation through a representational lens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14973–14987, Singapore. Association for Computational Linguistics.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024a. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024b. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Sandeep Subramanian, Oleksii Hrinchuk, Virginia Adams, and Oleksii Kuchaiev. 2021. NVIDIA NeMo's neural machine translation systems for English-German and English-Russian news and biomedical tasks at WMT21. In *Proceedings of the Sixth Conference on Machine Translation*, pages 197–204, Online. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

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.

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).

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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 4395–4405. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019b. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5797–5808. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George

Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Preprint*, arXiv:1609.08144.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. *Preprint*, arXiv:2309.17453.

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.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages. *Preprint*, arXiv:2305.18098.

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.

Shiyue Zhang, Vishrav Chaudhary, Naman Goyal, James Cross, Guillaume Wenzek, Mohit Bansal, and Francisco Guzman. 2022. How robust is neural machine translation to language imbalance in multilingual tokenizer training? In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 97–116, Orlando, USA. Association for Machine Translation in the Americas.

A Dataset

For each target language, we collected all openly licensed parallel data with Catalan from OPUS (Tiedemann, 2012). To augment dataset size, we also gathered parallel corpora with Spanish as the source language into Catalan using the Plan-TL Spanish-Catalan neural machine translation model[13], yielding synthetic Catalan corpora. These were concatenated with the original Catalan data and processed identically. In addition to OPUS data, we also used the Aina-ca-en-Parallel-Corpus for Catalan–Spanish pairs[14].

[13]Available on HuggingFace: https://huggingface.co/datasets/projecte-aina/CA-EN_Parallel_Corpus

[14]Available on HuggingFace: https://huggingface.co/PlanTL-GOB-ES/mt-plantl-es-ca

Dataset
Aina-ca-en-Parallel-Corpus
CCAligned
Covost2
DOGC
EUBookshop
Europarl
Globalvoices
Gnome
HLPT
KDE4
MultiCCAligned
NLLB
OpenSubtitles
ParaCrawl
Tatoeba
TildeModel
Ubuntu
Wikimatrix
Wikimedia
XLEnt

Table 8: Data sources.

Language	**Id**
Catalan	ca
German	de
English	en
Spanish	es
Basque	eu
Italian	it
Galician	gl
French	fr
Portuguese	pt

Table 9: List of BCP-47 language codes.

B Tokenizer

In our experiments, we utilized the BPE algorithm (Sennrich et al., 2016) from the *Huggingface Tokenizer* library (Moi and Patry, 2023). The settings used for training the tokenizer are detailed in Table 10. Every language tag is represented by a BCP-47 tag sequence where the base subtag is a three-letter

ISO 639-3 code, followed by ISO 15924 script subtags.

Hyper-Parameter	Value(s)
model_type	BPE
vocab_size	32k & 128k & 256k
nfkd_normalizer	True
lowercase_normalizer	False
pre_tokenizer	ByteLevel
add_prefix_space	False
special_tokens	<s>, </s>, <pad>, <mask>, [deu_Latn], [eng_Latn], [eus_Latn], [fra_Latn], [glg_Latn], [ita_Latn], [por_Latn], [spa_Latn], [cat_Latn]

Table 10: BPE tokenizer configuration.

We trained various tokenizers employing two distinct sampling strategies for each vocabulary size, then we evaluated them on fertility and parity (Petrov et al., 2024) metrics on FLORES-200 devtest. For a given tokenizer T and a set of sentences S, fertility is determined by dividing the total number of tokens generated from S (using T) by the total number of words in S. Parity is defined as achieving a balanced tokenization ratio between two languages. Specifically, a tokenizer T achieves parity for language A with respect to language B if the ratio $\frac{|T(s_A)|}{|T(s_B)|} \approx 1$, where s_A and s_B denote the sets of all sentences for languages A and B, respectively.

We experimented with both unigram and BPE implementations from the *Huggingface Tokenizer* library. We tested two sampling strategies: one involving the sampling of 1 million sentences from all languages, and another involving the equal sampling of 1 million sentences from Romance languages, with an oversampling of 3 million sentences for English, Basque, and German. Figure 7 presents the fertility metrics on English, Basque, and German. Given the results, we decided to use the BPE algorithm with the oversampling strategy for our final experiments. We also report obtained parity metrics by vocabulary size in figure 8.

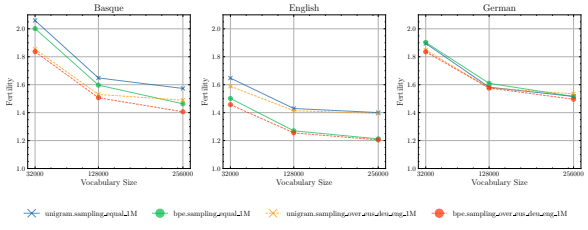


Figure 7: From left to right: fertility evaluated on Basque, English and German. Fertility is in the vertical axis, and vocabulary size is in the horizontal axis.

C Training

Hyper-Parameter	
Batch size	40
Number of Epochs	1
Optimizer	Adam
Adam-β_1	0.9
Adam-β_2	0.999
Adam-ϵ	1e-08
Learning rate	3e-04
LR Scheduler	Linear
Warmup Steps	2000

Table 11: Model training hyper-parameters

Num examples	26,301,993
Num tokens = Num examples * 2048 (considering pad tokens)	53,866,481,664
Num Epochs	1
Instantaneous batch size per device	1
Total train batch size (w. parallel, distributed & accumulation)	40
Gradient Accumulation steps	1
Total optimization steps	657,550
Number of trainable parameters	2,047,420,416

Table 12: Training and performance information for PLUME 32k.

Num examples	23,093,719
Num tokens = Num examples * 2048 (considering pad tokens)	47,295,936,512
Num Epochs	1
Instantaneous batch size per device	1
Total train batch size (w. parallel, distributed & accumulation)	40
Gradient Accumulation steps	1
Total optimization steps	577,343
Number of trainable parameters	2,244,028,416

Table 13: Training and performance information for PLUME 128k.

Num examples	22,213,825
Num tokens = Num examples * 2048 (considering pad tokens)	45,493,913,600
Num Epochs	1
Instantaneous batch size per device	1
Total train batch size (w. parallel, distributed & accumulation)	40
Gradient Accumulation steps	1
Total optimization steps	555,346
Number of trainable parameters	2,506,172,416

Table 14: Training and performance information for PLUME 256k.

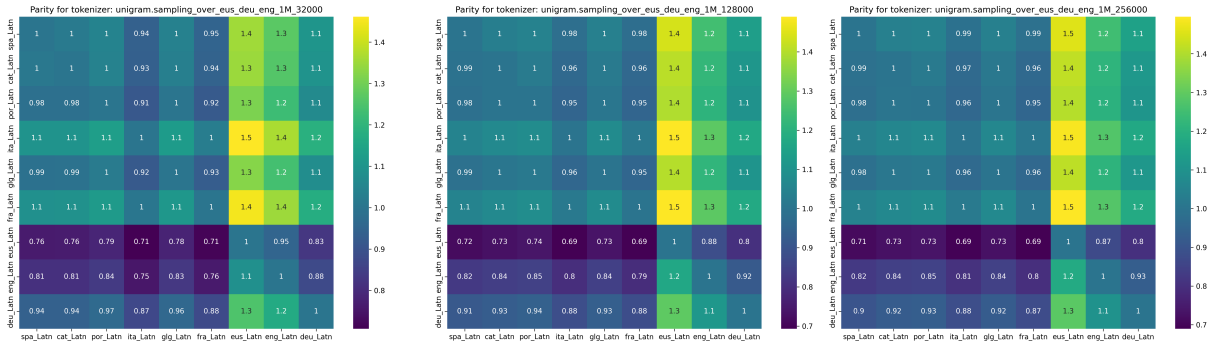


Figure 8: Parity for the different vocabulary sizes.

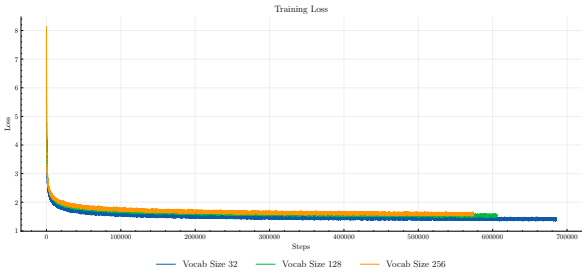


Figure 9: Training loss.

D Coverage metrics

We show in Figure 11 and Figure 12 the coverage heatmaps for PLUME 32k, 128k and 256k respectively. In Figure 13 we show the average coverage per layer for the different vocabulary sizes. We notice that PLUME 32k, 128k and 256k exhibit a similar coverage pattern across layers.

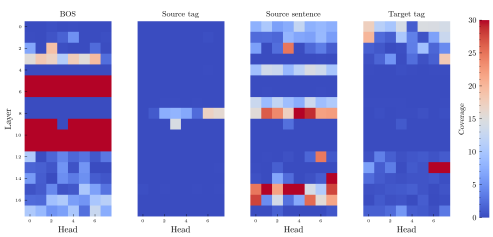


Figure 10: Coverage evaluating on FLORES-200 devtest using PLUME 32k. Each heatmap for each part of the prompt shows the coverage scores for each layer (vertical axis) and for each head (horizontal axis) in the model.

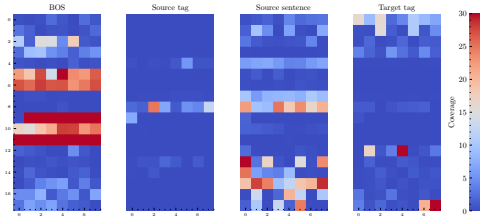


Figure 11: Coverage evaluating on FLORES-200 devtest using PLUME 128k. Each heatmap for each part of the prompt shows the coverage scores for each layer (vertical axis) and for each head (horizontal axis) in the model.

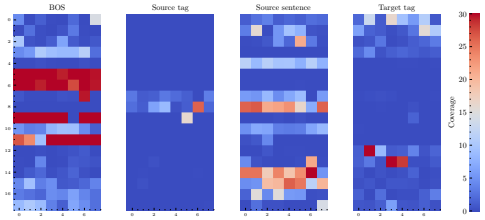


Figure 12: Coverage evaluating on FLORES-200 devtest using PLUME 256k. Each heatmap for each part of the prompt shows the coverage scores for each layer (vertical axis) and for each head (horizontal axis) in the model.

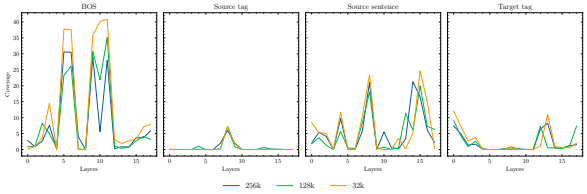


Figure 13: Average coverage per layer for each part of the prompt across various vocabulary sizes.

D.1 Attention matrices

An attention sink mechanism occurs when all the attention mass is given to some special tokens. We visualize the attention matrices for the first head of

layer 9 and layer 17 (last layer) in Figure 14. We observe that in layer 9, the model is giving all the attention mass to the BOS token[15] which allows the model to keep the residual stream of the network unchanged.

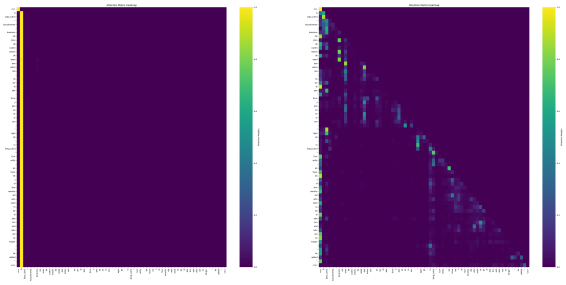


Figure 14: Attention weights for head 1 in layer 9 (left) and head 1 in layer 17 (right).

E Subspace distances

We show in Figure 15 the distances between language subspaces computed using the Riemannian metric on the space of positive definite matrices as detailed in (Chang et al., 2022) grouped by language and for each vocabulary size. We observe that for all the vocabulary sizes, Basque's subspace is further from the rest of the languages subspaces which could explain why model's performance on Basque is lower compared to other languages.

F UMAP Plots

Below we show the token representations[16] using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) for all the layers in PLUME 32k, 128k and 256k.

F.1 Spherical Voronoi diagrams

To better visualize high-dimensional token embeddings in PLUME models, we used spherical voronoi diagrams. Specifically, we reduced the embeddings to a 2D space, optimizing for cosine similarity using UMAP. Then, the 2D UMAP embeddings were projected onto a unit sphere. Specifically, each 2D point (x, y) was mapped to 3D coordinates (X, Y, Z) as follows:

$$X = \sin(x) \cdot \cos(y) \quad Y = \sin(x) \cdot \sin(y) \quad Z = \cos(x) \tag{4}$$

Then, for each language, we calculated the centroid of its corresponding tokens on the sphere and using these centroids, we computed Voronoi regions (where each region contains all the closest points to a specific centroid). We add as supplementary material the spherical voronoi diagrams for each layer in PLUME 32k.

G Detailed results

We report in the following tables the results of PLUME models for each translation direction. We also provide comparisons for TOWERBASE 7B (Alves et al., 2024) in those directions that PLUME and TOWERBASE 7B share, as well as comparisons with NLLB 3.3B (NLLB Team et al., 2022).

[15] There is a special token created by Huggingface BPE implementation, which is positioned between the BOS and the source tag tokens. We consider this special token as part of the BOS token.

[16] We use the first sentence from FLORES-200 devtest in each source language to construct the prompts: `"We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.`

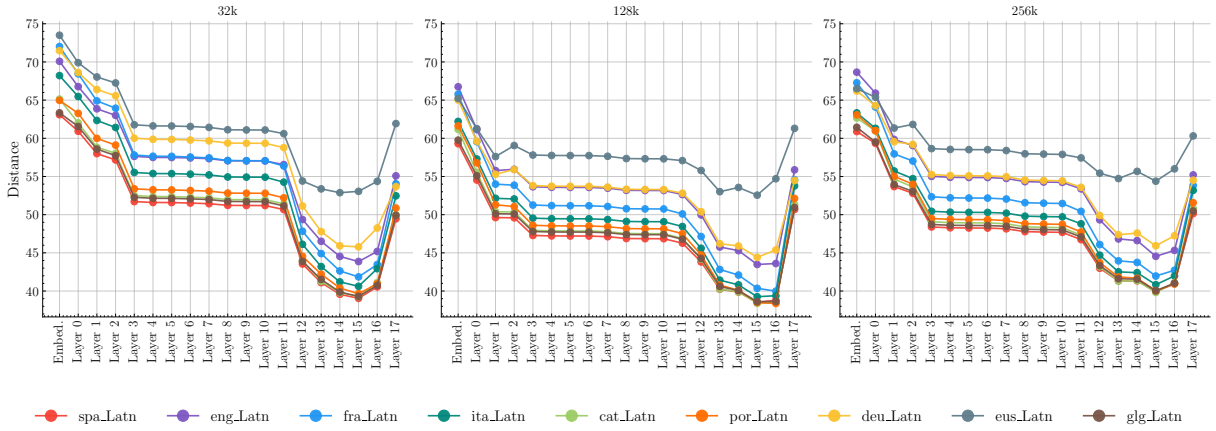


Figure 15: Mean distance between language subspaces grouped by languages and vocabulary sizes.

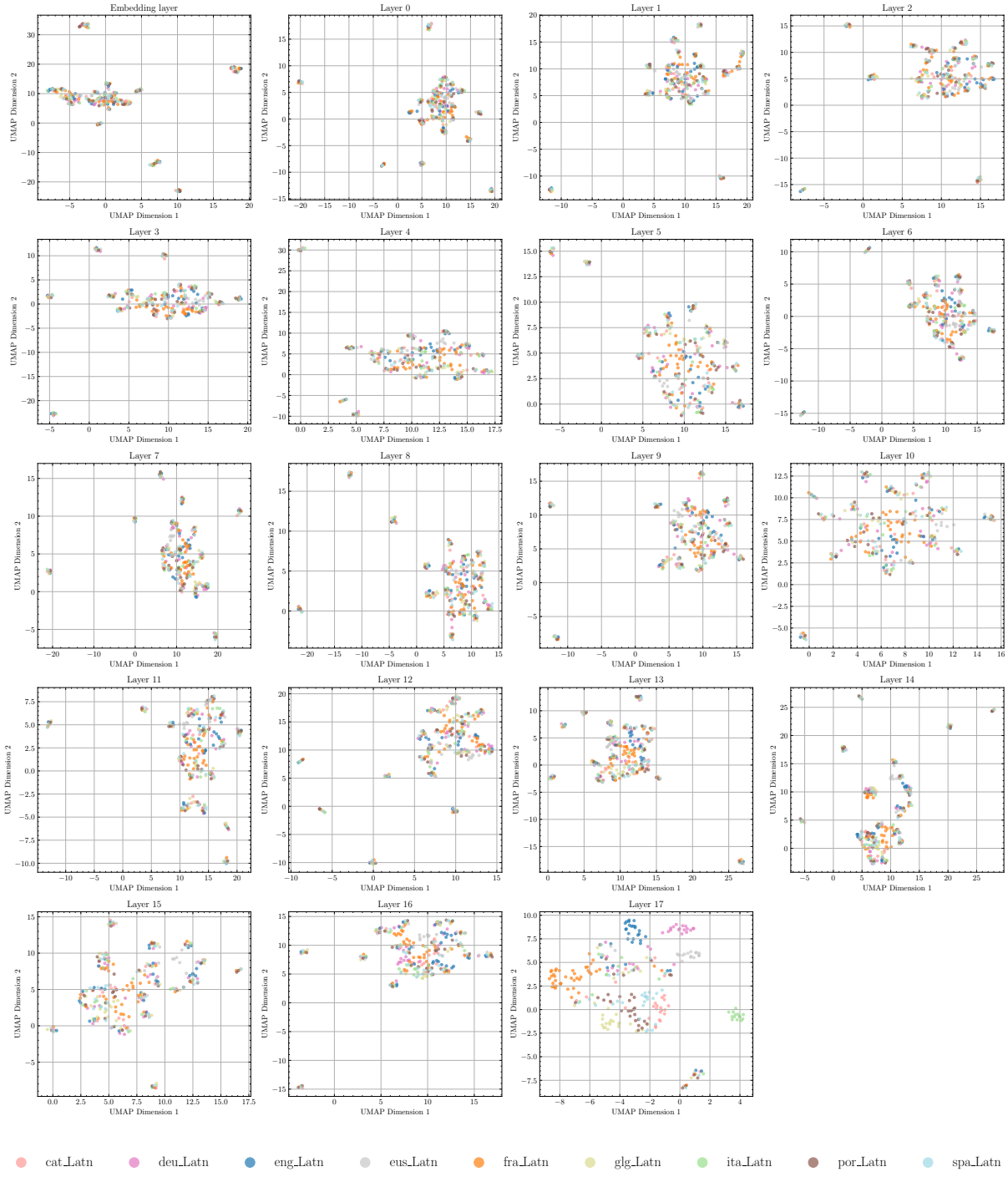


Figure 16: UMAP representations at the token embeddings in each layer grouped by source language using PLUME 32k.

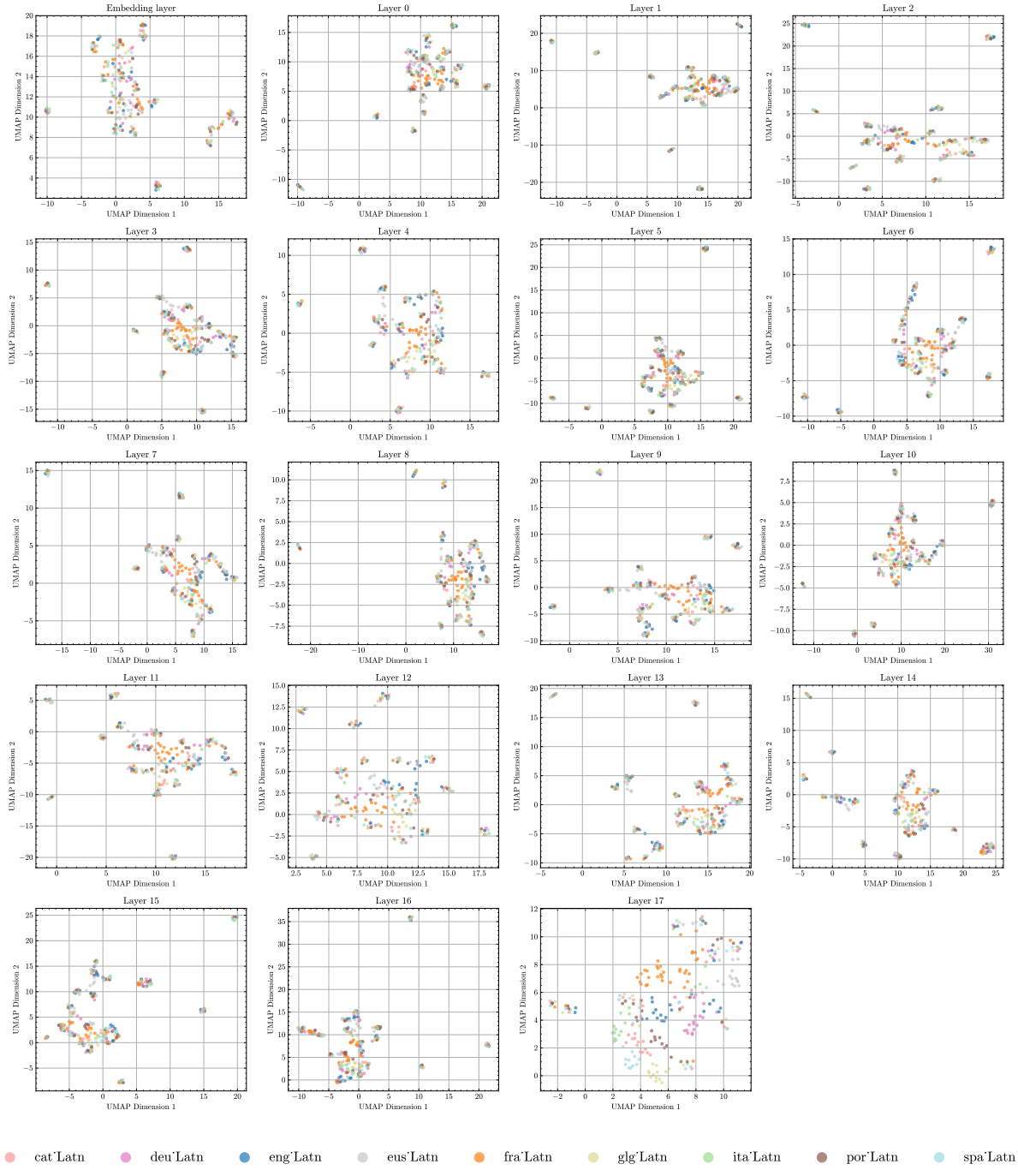


Figure 17: UMAP representations at the token embeddings in each layer grouped by source language using PLUME 128k.

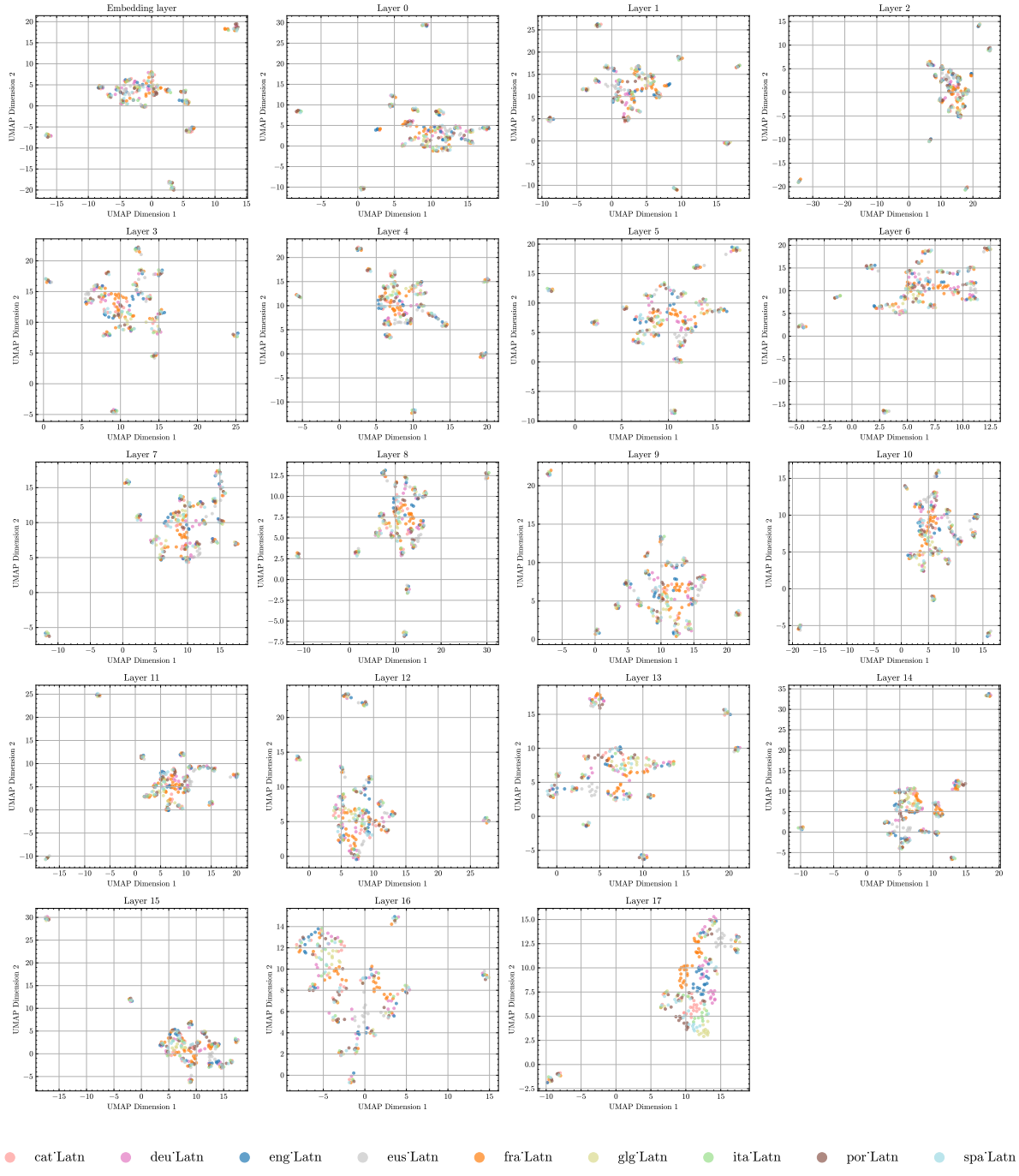


Figure 18: UMAP representations at the token embeddings in each layer grouped by source language using PLUME 256k.

Table 15: Results for ca→xx.

		FLORES-200				NTREX			
Pair	Model	BLEU	CHRF	COMET	COMET-KIWI	BLEU	CHRF	COMET	COMET-KIWI
ca-de	BSC Bilinguals	33.30	61.12	0.85	0.84	25.04	55.00	0.83	0.83
	NLLB 3.3B	31.19	58.41	0.85	0.84	21.72	53.41	0.81	0.82
	PLUME 128k	28.00	57.53	0.83	0.82	21.98	53.36	0.80	0.81
	PLUME 256k	28.55	57.63	0.83	0.82	21.39	52.72	0.80	0.81
	PLUME 32k	27.81	57.00	0.83	0.82	27.79	56.66	0.83	0.84
ca-en	BSC Bilinguals	46.29	70.44	0.88	0.86	41.20	66.57	0.87	0.86
	NLLB 3.3B	49.65	71.68	0.89	0.86	33.22	62.82	0.85	0.85
	PLUME 128k	42.91	68.69	0.88	0.86	33.73	63.07	0.85	0.85
	PLUME 256k	42.47	68.47	0.88	0.85	32.82	62.14	0.85	0.84
	PLUME 32k	41.92	68.15	0.87	0.85	37.61	64.98	0.87	0.85
ca-es	BSC Bilinguals	24.70	53.42	0.86	0.86	36.89	61.83	0.86	0.85
	NLLB 3.3B	25.62	53.73	0.86	0.86	35.44	61.27	0.86	0.85
	PLUME 128k	24.66	53.44	0.86	0.86	35.66	61.23	0.86	0.85
	PLUME 256k	24.59	53.37	0.86	0.85	35.70	61.24	0.86	0.85
	PLUME 32k	24.50	53.37	0.86	0.86	35.97	61.40	0.86	0.85
ca-eu	BSC Bilinguals	18.26	57.03	0.86	0.81	9.83	46.47	0.80	0.74
	NLLB 3.3B	13.13	50.47	0.83	0.75	12.40	49.99	0.82	0.78
	PLUME 128k	14.88	53.41	0.84	0.79	12.09	49.96	0.82	0.78
	PLUME 256k	14.97	53.75	0.84	0.78	12.17	49.58	0.81	0.77
	PLUME 32k	14.38	53.29	0.84	0.78	14.08	52.70	0.84	0.81
ca-fr	BSC Bilinguals	38.25	63.23	0.85	0.84	27.60	56.73	0.84	0.85
	NLLB 3.3B	39.89	64.05	0.86	0.85	25.20	54.13	0.81	0.82
	PLUME 128k	35.46	61.08	0.84	0.83	25.48	54.16	0.81	0.82
	PLUME 256k	35.72	61.18	0.84	0.83	24.94	53.76	0.81	0.82
	PLUME 32k	34.32	60.68	0.83	0.82	27.71	55.53	0.82	0.83
ca-gl	BSC Bilinguals	31.96	59.66	0.87	0.84	34.07	60.52	0.86	0.84
	NLLB 3.3B	32.78	59.25	0.87	0.85	33.23	60.22	0.86	0.84
	PLUME 128k	32.22	59.73	0.87	0.84	33.37	60.24	0.86	0.83
	PLUME 256k	32.07	59.51	0.87	0.84	33.23	60.27	0.86	0.84
	PLUME 32k	32.21	59.73	0.87	0.85	32.59	59.76	0.85	0.82
ca-it	BSC Bilinguals	26.92	56.55	0.87	0.85	29.46	58.00	0.87	0.85
	NLLB 3.3B	26.38	55.66	0.88	0.86	27.91	57.43	0.86	0.84
	PLUME 128k	25.77	55.78	0.87	0.85	28.11	57.62	0.86	0.84
	PLUME 256k	25.76	55.94	0.87	0.85	27.80	57.33	0.85	0.84
	PLUME 32k	25.45	55.51	0.87	0.85	29.07	57.95	0.86	0.84
ca-pt	BSC Bilinguals	37.18	62.73	0.88	0.84	31.46	57.67	0.86	0.84
	NLLB 3.3B	36.68	61.97	0.88	0.85	27.79	55.97	0.85	0.83
	PLUME 128k	36.27	62.12	0.88	0.84	28.50	56.29	0.85	0.83
	PLUME 256k	35.76	61.88	0.88	0.84	27.92	55.91	0.85	0.83
	PLUME 32k	35.81	61.67	0.88	0.84	28.19	56.17	0.85	0.83

Table 16: Results for de→xx.

		FLORES-200				NTREX			
Pair	Model	BLEU	CHRF	COMET	COMET-KIWI	BLEU	CHRF	COMET	COMET-KIWI
de-ca	BSC Bilinguals	30.15	57.65	0.83	0.82	28.24	55.02	0.83	0.84
	NLLB 3.3B	31.45	57.99	0.86	0.85	28.34	55.03	0.82	0.82
	PLUME 128k	32.23	59.02	0.85	0.83	28.13	54.66	0.82	0.82
	PLUME 256k	31.76	58.73	0.85	0.83	27.94	54.58	0.82	0.81
	PLUME 32k	31.76	58.56	0.85	0.83	24.49	53.60	0.78	0.80
de-en	NLLB 3.3B	46.02	69.30	0.90	0.85	41.01	66.16	0.88	0.84
	TOWERBASE 7B	43.69	68.56	0.89	0.84	41.01	66.16	0.88	0.84
	PLUME 128k	36.17	63.49	0.86	0.82	29.73	59.26	0.84	0.81
	PLUME 256k	36.99	64.04	0.87	0.83	29.80	59.39	0.84	0.81
	PLUME 32k	34.12	62.13	0.86	0.81	28.73	58.11	0.83	0.80
de-es	NLLB 3.3B	23.86	51.39	0.84	0.86	31.13	57.36	0.84	0.85
	TOWERBASE 7B	21.66	50.94	0.83	0.85	31.13	57.36	0.84	0.85
	PLUME 128k	22.00	50.41	0.82	0.83	28.41	54.92	0.81	0.82
	PLUME 256k	22.35	50.80	0.82	0.83	28.76	54.89	0.81	0.82
	PLUME 32k	20.90	49.74	0.82	0.82	27.83	54.18	0.81	0.81
de-eu	NLLB 3.3B	9.83	45.23	0.78	0.71	7.83	41.70	0.76	0.69
	PLUME 128k	9.91	46.23	0.78	0.73	8.18	42.65	0.75	0.72
	PLUME 256k	11.48	47.52	0.79	0.74	8.93	43.59	0.76	0.73
	PLUME 32k	10.77	46.22	0.77	0.72	8.46	42.39	0.74	0.71
de-fr	NLLB 3.3B	37.62	62.60	0.86	0.85	28.06	56.03	0.83	0.85
	TOWERBASE 7B	34.84	61.23	0.85	0.85	28.06	56.03	0.83	0.85
	PLUME 128k	28.50	56.32	0.80	0.80	20.26	49.16	0.77	0.78
	PLUME 256k	29.01	56.15	0.80	0.79	20.84	49.13	0.77	0.78
	PLUME 32k	27.13	54.89	0.79	0.78	20.37	48.30	0.75	0.76
de-gl	NLLB 3.3B	28.87	55.70	0.85	0.85	29.17	56.21	0.84	0.84
	PLUME 128k	26.01	54.15	0.83	0.83	24.55	52.87	0.81	0.81
	PLUME 256k	25.20	53.46	0.83	0.82	24.87	52.86	0.81	0.81
	PLUME 32k	25.31	53.11	0.82	0.82	24.11	51.92	0.80	0.80
de-it	NLLB 3.3B	25.88	54.95	0.87	0.86	27.84	56.12	0.86	0.85
	TOWERBASE 7B	24.73	54.26	0.86	0.85	27.84	56.12	0.86	0.85
	PLUME 128k	22.47	52.44	0.84	0.83	22.77	52.04	0.82	0.82
	PLUME 256k	22.74	52.34	0.85	0.83	23.12	52.16	0.82	0.82
	PLUME 32k	21.36	51.19	0.84	0.82	22.39	51.53	0.81	0.81
de-pt	NLLB 3.3B	33.42	59.32	0.87	0.85	29.42	55.97	0.85	0.85
	TOWERBASE 7B	30.94	58.48	0.86	0.85	29.42	55.97	0.85	0.85
	PLUME 128k	30.02	57.17	0.85	0.83	24.09	51.90	0.82	0.82
	PLUME 256k	30.36	57.46	0.85	0.83	24.06	51.90	0.82	0.82
	PLUME 32k	29.19	55.98	0.84	0.81	23.00	51.09	0.80	0.80

Table 17: Results for en→xx.

		FLORES-200				NTREX			
Pair	Model	BLEU	CHRF	COMET	COMET-KIWI	BLEU	CHRF	COMET	COMET-KIWI
en-ca	BSC Bilinguals	44.05	67.95	0.88	0.85	37.49	62.38	0.85	0.83
	NLLB 3.3B	42.33	65.97	0.88	0.85	35.80	61.29	0.83	0.81
	PLUME 128k	42.29	66.44	0.87	0.84	35.95	61.30	0.83	0.81
	PLUME 256k	42.64	66.59	0.87	0.84	35.05	60.72	0.82	0.80
	PLUME 32k	42.32	66.39	0.86	0.84	37.93	63.19	0.84	0.82
en-de	NLLB 3.3B	39.88	65.14	0.88	0.84	32.46	60.93	0.85	0.84
	TOWERBASE 7B	37.53	64.47	0.87	0.84	32.46	60.93	0.85	0.84
	PLUME 128k	31.27	59.30	0.82	0.80	24.31	54.33	0.78	0.77
	PLUME 256k	31.81	60.17	0.83	0.81	24.94	55.13	0.79	0.78
	PLUME 32k	29.86	58.22	0.82	0.79	23.46	53.42	0.77	0.75
en-es	NLLB 3.3B	28.14	55.85	0.86	0.86	39.33	63.79	0.85	0.84
	TOWERBASE 7B	26.38	55.02	0.86	0.86	39.33	63.79	0.85	0.84
	PLUME 128k	24.34	53.01	0.83	0.84	35.62	60.75	0.81	0.80
	PLUME 256k	25.00	53.43	0.84	0.84	36.42	61.36	0.82	0.81
	PLUME 32k	23.47	52.61	0.83	0.83	34.86	60.10	0.81	0.79
en-eu	NLLB 3.3B	15.71	53.25	0.85	0.82	11.62	47.74	0.81	0.79
	PLUME 128k	13.02	48.69	0.81	0.78	10.51	44.21	0.76	0.75
	PLUME 256k	12.95	50.05	0.81	0.79	10.96	45.41	0.77	0.75
	PLUME 32k	13.03	48.89	0.80	0.78	10.73	44.79	0.75	0.74
en-fr	NLLB 3.3B	50.90	71.70	0.88	0.87	34.77	61.69	0.84	0.85
	TOWERBASE 7B	49.28	70.83	0.88	0.87	34.77	61.69	0.84	0.85
	PLUME 128k	36.49	62.25	0.82	0.82	26.36	54.27	0.77	0.79
	PLUME 256k	38.27	63.03	0.83	0.83	27.20	54.95	0.77	0.79
	PLUME 32k	36.11	61.92	0.81	0.81	26.36	54.15	0.76	0.78
en-gl	NLLB 3.3B	35.98	61.55	0.87	0.85	39.01	63.75	0.85	0.83
	PLUME 128k	32.26	59.64	0.85	0.83	33.28	59.53	0.81	0.79
	PLUME 256k	32.61	59.66	0.85	0.83	33.13	59.59	0.81	0.79
	PLUME 32k	31.16	58.92	0.84	0.82	31.88	58.48	0.80	0.77
en-it	NLLB 3.3B	30.63	59.52	0.88	0.87	37.68	63.84	0.87	0.85
	TOWERBASE 7B	29.64	59.13	0.88	0.87	37.68	63.84	0.87	0.85
	PLUME 128k	25.58	55.15	0.84	0.84	28.84	57.37	0.82	0.81
	PLUME 256k	25.64	55.75	0.85	0.85	30.73	58.42	0.82	0.81
	PLUME 32k	24.51	54.69	0.84	0.84	29.55	57.32	0.81	0.80
en-pt	NLLB 3.3B	49.45	70.54	0.90	0.85	37.37	62.46	0.87	0.84
	TOWERBASE 7B	49.67	71.36	0.90	0.85	37.37	62.46	0.87	0.84
	PLUME 128k	40.94	65.75	0.87	0.83	30.59	57.41	0.82	0.79
	PLUME 256k	42.62	66.47	0.87	0.83	31.27	57.81	0.82	0.79
	PLUME 32k	40.57	65.13	0.86	0.82	30.13	56.87	0.81	0.78

Table 18: Results for es→xx.

		FLORES-200				NTREX			
Pair	Model	BLEU	CHRF	COMET	COMET-KIWI	BLEU	CHRF	COMET	COMET-KIWI
es-ca	BSC Bilinguals	23.34	53.98	0.86	0.84	34.47	60.52	0.86	0.84
	NLLB 3.3B	25.70	55.24	0.86	0.84	33.16	60.59	0.86	0.83
	PLUME 128k	23.43	54.22	0.86	0.84	33.41	60.49	0.86	0.83
	PLUME 256k	23.42	54.20	0.86	0.84	33.23	60.60	0.86	0.83
	PLUME 32k	23.55	54.30	0.86	0.84	34.14	60.73	0.86	0.83
es-de	NLLB 3.3B	22.88	53.27	0.84	0.84	24.63	55.15	0.83	0.84
	TOWERBASE 7B	18.86	51.44	0.82	0.84	24.63	55.15	0.83	0.84
	PLUME 128k	17.69	50.73	0.80	0.81	19.90	52.08	0.79	0.81
	PLUME 256k	18.06	51.26	0.81	0.82	20.41	52.30	0.80	0.81
	PLUME 32k	17.63	50.19	0.80	0.80	19.47	51.49	0.78	0.80
es-en	NLLB 3.3B	32.93	61.52	0.88	0.86	41.88	67.47	0.88	0.86
	TOWERBASE 7B	30.47	60.37	0.87	0.86	41.88	67.47	0.88	0.86
	PLUME 128k	24.74	56.76	0.85	0.85	31.64	62.07	0.85	0.84
	PLUME 256k	24.91	57.16	0.85	0.85	31.53	62.24	0.85	0.84
	PLUME 32k	23.79	56.29	0.84	0.85	31.05	61.38	0.85	0.84
es-eu	NLLB 3.3B	11.31	49.93	0.84	0.81	11.13	47.56	0.81	0.77
	PLUME 128k	10.39	49.12	0.82	0.81	11.45	48.54	0.81	0.79
	PLUME 256k	11.22	49.59	0.83	0.81	11.29	48.92	0.81	0.79
	PLUME 32k	11.26	49.16	0.82	0.79	11.31	47.79	0.80	0.78
es-fr	NLLB 3.3B	29.97	58.18	0.85	0.86	27.92	56.77	0.84	0.85
	TOWERBASE 7B	25.16	55.84	0.84	0.85	27.92	56.77	0.84	0.85
	PLUME 128k	21.91	52.76	0.81	0.82	23.99	52.86	0.80	0.81
	PLUME 256k	22.15	52.87	0.81	0.82	23.85	52.99	0.80	0.81
	PLUME 32k	21.96	52.78	0.81	0.82	24.39	53.10	0.79	0.81
es-gl	NLLB 3.3B	24.64	53.77	0.87	0.84	34.92	61.24	0.87	0.83
	PLUME 128k	21.47	52.69	0.87	0.84	33.34	60.71	0.86	0.83
	PLUME 256k	21.59	52.54	0.86	0.84	33.63	60.81	0.86	0.82
	PLUME 32k	21.29	52.51	0.86	0.84	33.08	60.63	0.86	0.83
es-it	NLLB 3.3B	22.77	52.86	0.87	0.86	29.60	58.19	0.87	0.85
	TOWERBASE 7B	19.95	51.18	0.86	0.86	29.60	58.19	0.87	0.85
	PLUME 128k	18.76	50.27	0.85	0.85	25.08	55.31	0.84	0.83
	PLUME 256k	18.86	50.53	0.85	0.84	25.42	55.57	0.85	0.84
	PLUME 32k	19.29	50.45	0.85	0.84	25.14	55.55	0.84	0.83
es-pt	NLLB 3.3B	26.18	55.23	0.87	0.85	32.30	58.24	0.87	0.84
	TOWERBASE 7B	23.11	53.87	0.87	0.85	32.30	58.24	0.87	0.84
	PLUME 128k	21.16	52.25	0.86	0.84	25.82	54.84	0.85	0.83
	PLUME 256k	21.84	52.70	0.86	0.84	27.27	55.53	0.85	0.83
	PLUME 32k	21.65	52.74	0.86	0.84	27.00	55.35	0.85	0.83

Table 19: Results for eu→xx.

		FLORES-200				NTREX			
Pair	Model	BLEU	CHRF	COMET	COMET-KIWI	BLEU	CHRF	COMET	COMET-KIWI
eu-ca	BSC Bilinguals	26.18	54.14	0.85	0.82	24.56	51.56	0.83	0.81
	NLLB 3.3B	26.70	53.97	0.86	0.82	22.29	49.79	0.81	0.79
	PLUME 128k	24.33	51.85	0.84	0.80	21.70	49.48	0.81	0.78
	PLUME 256k	24.02	51.67	0.84	0.80	20.19	48.69	0.80	0.77
	PLUME 32k	22.92	50.69	0.83	0.79	24.29	51.84	0.82	0.81
eu-de	NLLB 3.3B	22.71	51.75	0.83	0.80	18.96	48.84	0.81	0.79
	PLUME 128k	13.64	44.72	0.76	0.72	11.38	41.74	0.73	0.72
	PLUME 256k	13.58	44.77	0.76	0.72	10.74	41.78	0.73	0.72
	PLUME 32k	10.62	40.74	0.72	0.69	9.30	38.93	0.69	0.69
eu-en	NLLB 3.3B	33.44	60.57	0.87	0.86	29.59	57.37	0.85	0.85
	PLUME 128k	21.49	51.65	0.82	0.81	16.70	48.58	0.79	0.80
	PLUME 256k	22.12	52.31	0.82	0.82	16.41	48.54	0.79	0.80
	PLUME 32k	17.52	48.60	0.79	0.78	13.84	45.54	0.77	0.77
eu-es	NLLB 3.3B	20.50	48.29	0.84	0.84	27.50	53.84	0.84	0.83
	PLUME 128k	17.74	45.98	0.81	0.81	20.71	48.75	0.79	0.79
	PLUME 256k	17.94	45.41	0.81	0.81	20.58	48.54	0.79	0.79
	PLUME 32k	15.61	43.47	0.79	0.79	18.76	47.03	0.78	0.78
eu-fr	NLLB 3.3B	29.05	56.00	0.84	0.83	22.63	50.58	0.81	0.82
	PLUME 128k	18.58	46.77	0.75	0.75	14.90	42.94	0.73	0.73
	PLUME 256k	18.39	46.08	0.75	0.74	14.73	42.58	0.72	0.72
	PLUME 32k	15.77	44.00	0.71	0.71	12.58	40.59	0.69	0.70
eu-gl	NLLB 3.3B	25.16	52.52	0.86	0.83	24.18	52.15	0.83	0.82
	PLUME 128k	19.24	47.58	0.82	0.78	18.04	46.91	0.79	0.77
	PLUME 256k	18.53	46.92	0.81	0.78	18.23	46.74	0.79	0.76
	PLUME 32k	15.91	45.11	0.79	0.75	16.13	44.99	0.77	0.75
eu-it	NLLB 3.3B	21.27	51.07	0.86	0.84	22.45	51.13	0.84	0.83
	PLUME 128k	16.39	45.65	0.81	0.80	16.82	46.45	0.79	0.79
	PLUME 256k	16.46	45.76	0.81	0.80	15.96	46.05	0.79	0.78
	PLUME 32k	14.01	43.52	0.79	0.77	14.34	44.19	0.77	0.76
eu-pt	NLLB 3.3B	27.79	54.65	0.86	0.84	23.93	50.72	0.83	0.82
	PLUME 128k	20.12	48.58	0.82	0.80	16.11	44.79	0.79	0.78
	PLUME 256k	20.89	48.87	0.81	0.80	16.80	45.27	0.79	0.78
	PLUME 32k	17.64	46.34	0.79	0.77	14.05	42.96	0.76	0.76

Table 20: Results for fr→xx.

		FLORES-200				NTREX			
Pair	Model	BLEU	CHRF	COMET	COMET-KIWI	BLEU	CHRF	COMET	COMET-KIWI
fr-ca	BSC Bilinguals	34.44	60.10	0.86	0.83	29.22	55.76	0.84	0.83
	NLLB 3.3B	34.00	59.82	0.87	0.84	27.30	54.40	0.83	0.82
	PLUME 128k	34.35	60.24	0.86	0.83	27.57	54.40	0.83	0.81
	PLUME 256k	33.63	59.83	0.86	0.83	27.00	54.18	0.83	0.81
	PLUME 32k	34.28	60.16	0.86	0.83	27.03	54.04	0.83	0.81
fr-de	NLLB 3.3B	29.96	57.73	0.85	0.84	23.82	53.55	0.83	0.84
	TOWERBASE 7B	25.48	56.02	0.82	0.84	23.82	53.55	0.83	0.84
	PLUME 128k	24.63	54.96	0.81	0.80	19.07	49.59	0.78	0.78
	PLUME 256k	23.85	54.54	0.82	0.80	18.18	49.18	0.78	0.78
	PLUME 32k	22.45	53.56	0.81	0.78	18.35	48.80	0.77	0.77
fr-en	NLLB 3.3B	48.38	70.72	0.90	0.86	40.30	64.78	0.87	0.86
	TOWERBASE 7B	45.48	69.54	0.89	0.86	40.30	64.78	0.87	0.86
	PLUME 128k	37.37	64.47	0.87	0.85	28.95	58.15	0.84	0.84
	PLUME 256k	37.74	64.80	0.87	0.85	29.11	58.37	0.84	0.84
	PLUME 32k	34.87	63.11	0.86	0.84	28.36	57.38	0.83	0.83
fr-es	NLLB 3.3B	24.45	52.39	0.86	0.86	32.28	57.85	0.85	0.85
	TOWERBASE 7B	22.02	51.42	0.84	0.85	32.28	57.85	0.85	0.85
	PLUME 128k	21.65	50.63	0.84	0.84	27.18	54.18	0.82	0.83
	PLUME 256k	21.80	50.74	0.84	0.84	27.30	54.22	0.82	0.83
	PLUME 32k	21.60	50.66	0.84	0.84	27.23	54.00	0.82	0.82
fr-eu	NLLB 3.3B	10.73	46.16	0.80	0.73	7.79	41.10	0.76	0.69
	PLUME 128k	10.79	48.17	0.80	0.76	9.32	44.51	0.78	0.75
	PLUME 256k	11.78	48.71	0.80	0.77	9.43	44.37	0.78	0.75
	PLUME 32k	11.59	48.08	0.79	0.75	8.65	43.30	0.76	0.72
fr-gl	NLLB 3.3B	30.59	57.45	0.86	0.85	29.61	56.42	0.85	0.84
	PLUME 128k	27.95	55.92	0.85	0.84	24.65	52.84	0.81	0.81
	PLUME 256k	28.49	55.94	0.85	0.84	24.57	52.94	0.82	0.81
	PLUME 32k	27.69	55.65	0.85	0.83	24.11	52.42	0.81	0.81
fr-it	NLLB 3.3B	27.06	56.27	0.88	0.86	28.22	56.47	0.86	0.86
	TOWERBASE 7B	25.14	55.00	0.87	0.86	28.22	56.47	0.86	0.86
	PLUME 128k	24.45	53.92	0.86	0.84	24.25	53.18	0.84	0.83
	PLUME 256k	24.27	53.92	0.86	0.84	24.45	53.22	0.84	0.83
	PLUME 32k	23.98	53.72	0.86	0.84	23.84	53.05	0.83	0.82
fr-pt	NLLB 3.3B	36.18	61.28	0.88	0.85	29.11	55.64	0.85	0.84
	TOWERBASE 7B	33.03	60.10	0.87	0.85	29.11	55.64	0.85	0.84
	PLUME 128k	32.15	59.00	0.86	0.83	24.59	52.51	0.83	0.82
	PLUME 256k	32.86	59.22	0.86	0.83	24.85	52.21	0.82	0.81
	PLUME 32k	31.72	58.70	0.86	0.82	24.33	52.19	0.82	0.81

Table 21: Results for it→xx.

		FLORES-200				NTREX			
Pair	Model	BLEU	CHRF	COMET	COMET-KIWI	BLEU	CHRF	COMET	COMET-KIWI
it-ca	BSC Bilinguals	27.68	56.63	0.86	0.84	31.87	57.96	0.86	0.84
	NLLB 3.3B	27.77	56.56	0.87	0.86	31.18	57.64	0.85	0.83
	PLUME 128k	27.92	57.34	0.87	0.85	31.00	57.62	0.85	0.83
	PLUME 256k	27.86	57.25	0.87	0.85	30.69	57.35	0.85	0.83
	PLUME 32k	27.48	57.19	0.86	0.85	30.67	57.08	0.84	0.82
it-de	NLLB 3.3B	25.33	55.23	0.85	0.86	26.76	56.82	0.84	0.85
	TOWERBASE 7B	18.14	49.13	0.82	0.86	26.76	56.82	0.84	0.85
	PLUME 128k	20.84	52.75	0.82	0.83	20.84	51.69	0.79	0.82
	PLUME 256k	21.05	53.04	0.82	0.83	21.06	52.07	0.80	0.82
	PLUME 32k	19.77	51.78	0.81	0.82	20.28	51.35	0.79	0.80
it-en	NLLB 3.3B	36.33	64.25	0.88	0.87	43.96	67.59	0.88	0.86
	TOWERBASE 7B	32.95	62.57	0.88	0.86	43.96	67.59	0.88	0.86
	PLUME 128k	27.80	58.98	0.86	0.85	33.76	62.30	0.85	0.84
	PLUME 256k	28.91	59.82	0.86	0.86	34.76	62.75	0.85	0.85
	PLUME 32k	27.43	58.75	0.85	0.85	32.90	61.49	0.84	0.84
it-es	NLLB 3.3B	22.70	51.45	0.86	0.87	34.15	59.45	0.86	0.86
	TOWERBASE 7B	20.71	50.87	0.85	0.87	34.15	59.45	0.86	0.86
	PLUME 128k	20.91	50.70	0.85	0.86	30.30	56.88	0.84	0.85
	PLUME 256k	21.35	51.04	0.85	0.86	30.62	56.96	0.84	0.85
	PLUME 32k	20.99	50.72	0.85	0.86	30.06	56.70	0.84	0.85
it-eu	NLLB 3.3B	7.65	43.50	0.79	0.73	8.09	41.63	0.76	0.70
	PLUME 128k	9.77	47.74	0.81	0.79	10.07	45.74	0.79	0.76
	PLUME 256k	11.33	49.20	0.82	0.80	10.82	46.47	0.79	0.77
	PLUME 32k	10.69	48.55	0.81	0.78	10.44	45.82	0.78	0.76
it-fr	NLLB 3.3B	33.24	60.44	0.87	0.87	29.23	57.43	0.84	0.86
	TOWERBASE 7B	29.16	58.49	0.85	0.87	29.23	57.43	0.84	0.86
	PLUME 128k	27.21	56.24	0.83	0.84	23.92	52.66	0.81	0.82
	PLUME 256k	27.89	56.11	0.83	0.84	24.39	52.83	0.80	0.82
	PLUME 32k	26.35	55.67	0.82	0.83	24.04	52.53	0.80	0.81
it-gl	NLLB 3.3B	25.72	54.62	0.87	0.86	32.39	58.86	0.86	0.84
	PLUME 128k	23.80	54.06	0.86	0.85	29.04	56.66	0.84	0.83
	PLUME 256k	23.79	53.94	0.86	0.84	29.34	56.60	0.84	0.82
	PLUME 32k	23.59	53.88	0.85	0.84	28.20	55.97	0.84	0.82
it-pt	NLLB 3.3B	28.17	56.94	0.88	0.86	33.41	58.86	0.87	0.85
	TOWERBASE 7B	24.49	55.37	0.86	0.85	33.41	58.86	0.87	0.85
	PLUME 128k	26.64	56.24	0.87	0.84	28.48	55.43	0.85	0.83
	PLUME 256k	27.10	56.52	0.87	0.85	28.33	55.31	0.84	0.83
	PLUME 32k	25.86	55.58	0.86	0.84	28.03	55.24	0.84	0.82

Table 22: Results for gl→xx.

		FLORES-200				NTREX			
Pair	Model	BLEU	CHRF	COMET	COMET-KIWI	BLEU	CHRF	COMET	COMET-KIWI
gl-ca	BSC Bilinguals	32.43	60.50	0.87	0.84	34.27	60.27	0.86	0.84
	NLLB 3.3B	34.43	60.88	0.87	0.85	34.25	60.34	0.86	0.83
	PLUME 128k	32.77	60.71	0.87	0.84	34.28	60.55	0.86	0.83
	PLUME 256k	33.00	60.85	0.88	0.84	34.10	60.42	0.86	0.83
	PLUME 32k	32.75	60.76	0.87	0.84	33.72	60.27	0.86	0.83
gl-de	NLLB 3.3B	29.57	57.53	0.85	0.84	25.13	55.12	0.83	0.83
	PLUME 128k	23.05	54.44	0.81	0.81	20.23	51.72	0.79	0.80
	PLUME 256k	24.25	55.47	0.82	0.82	20.35	52.31	0.79	0.80
	PLUME 32k	22.89	54.11	0.80	0.80	19.75	51.46	0.78	0.79
gl-en	NLLB 3.3B	44.14	68.60	0.89	0.86	43.52	67.80	0.88	0.85
	PLUME 128k	35.47	64.50	0.86	0.85	33.40	62.42	0.85	0.84
	PLUME 256k	34.74	64.17	0.86	0.84	32.56	62.21	0.85	0.84
	PLUME 32k	34.15	63.48	0.86	0.84	30.76	61.22	0.84	0.83
gl-es	NLLB 3.3B	25.59	53.47	0.87	0.85	36.99	61.92	0.87	0.84
	PLUME 128k	23.67	52.86	0.86	0.85	35.18	61.04	0.86	0.84
	PLUME 256k	23.79	52.87	0.86	0.85	35.84	61.32	0.86	0.84
	PLUME 32k	23.59	52.83	0.86	0.85	35.48	61.15	0.86	0.84
gl-eu	NLLB 3.3B	12.37	48.45	0.82	0.73	9.06	43.94	0.78	0.70
	PLUME 128k	13.23	51.10	0.83	0.77	11.89	48.13	0.81	0.76
	PLUME 256k	13.68	51.27	0.83	0.77	11.28	48.44	0.81	0.76
	PLUME 32k	12.78	50.05	0.82	0.75	10.94	47.31	0.80	0.74
gl-fr	NLLB 3.3B	38.37	63.38	0.86	0.85	29.03	56.98	0.84	0.84
	PLUME 128k	29.14	57.49	0.82	0.82	23.19	52.26	0.79	0.81
	PLUME 256k	30.24	57.82	0.82	0.82	23.80	52.55	0.79	0.80
	PLUME 32k	29.84	57.65	0.81	0.81	23.56	52.22	0.79	0.80
gl-it	NLLB 3.3B	26.14	55.52	0.88	0.85	30.79	58.39	0.87	0.84
	PLUME 128k	22.73	53.29	0.86	0.84	26.47	55.68	0.84	0.83
	PLUME 256k	23.20	53.77	0.86	0.84	27.00	56.19	0.84	0.83
	PLUME 32k	22.45	53.22	0.86	0.84	26.36	55.84	0.84	0.83
gl-pt	NLLB 3.3B	34.42	60.37	0.88	0.83	31.87	58.16	0.87	0.83
	PLUME 128k	28.42	57.24	0.87	0.83	26.36	54.81	0.85	0.81
	PLUME 256k	29.11	57.70	0.87	0.83	27.82	55.65	0.85	0.81
	PLUME 32k	29.23	57.83	0.87	0.83	27.50	55.41	0.85	0.81

Table 23: Results for pt→xx.

		FLORES-200				NTREX			
Pair	Model	BLEU	CHRF	COMET	COMET-KIWI	BLEU	CHRF	COMET	COMET-KIWI
pt-ca	BSC Bilinguals	35.75	61.22	0.87	0.84	32.04	58.28	0.86	0.83
	NLLB 3.3B	34.64	60.68	0.87	0.84	31.17	57.91	0.85	0.83
	PLUME 128k	35.50	61.41	0.87	0.84	31.05	57.84	0.85	0.83
	PLUME 256k	35.38	60.95	0.87	0.83	31.12	57.84	0.85	0.83
	PLUME 32k	35.50	61.26	0.87	0.83	30.95	57.66	0.85	0.82
pt-de	NLLB 3.3B	31.27	58.75	0.85	0.85	25.56	55.62	0.84	0.84
	TOWERBASE 7B	25.48	56.02	0.82	0.84	25.56	55.62	0.84	0.84
	PLUME 128k	25.45	55.44	0.82	0.82	19.99	51.73	0.80	0.80
	PLUME 256k	26.51	55.90	0.83	0.82	20.03	51.96	0.80	0.81
	PLUME 32k	25.01	54.48	0.81	0.81	20.48	51.29	0.79	0.79
pt-en	NLLB 3.3B	52.50	73.31	0.90	0.85	43.94	68.11	0.88	0.85
	TOWERBASE 7B	50.16	72.76	0.90	0.85	43.94	68.11	0.88	0.85
	PLUME 128k	42.71	68.42	0.88	0.84	33.21	62.26	0.85	0.83
	PLUME 256k	43.31	68.95	0.88	0.84	33.50	62.46	0.86	0.83
	PLUME 32k	41.73	67.58	0.87	0.83	32.87	61.63	0.85	0.82
pt-es	NLLB 3.3B	25.76	53.31	0.86	0.86	34.85	60.45	0.86	0.85
	TOWERBASE 7B	22.82	51.90	0.85	0.85	34.85	60.45	0.86	0.85
	PLUME 128k	22.97	51.85	0.85	0.85	30.89	57.40	0.85	0.84
	PLUME 256k	23.04	51.82	0.85	0.84	31.32	57.66	0.85	0.84
	PLUME 32k	22.72	51.74	0.85	0.84	30.84	57.25	0.85	0.84
pt-eu	NLLB 3.3B	10.38	45.45	0.79	0.72	8.14	41.30	0.76	0.69
	PLUME 128k	11.18	49.09	0.82	0.79	9.93	46.18	0.80	0.77
	PLUME 256k	13.37	50.70	0.82	0.79	10.26	46.86	0.80	0.77
	PLUME 32k	12.68	49.77	0.81	0.78	10.50	46.72	0.79	0.76
pt-fr	NLLB 3.3B	40.85	64.94	0.87	0.86	29.39	57.41	0.84	0.85
	TOWERBASE 7B	36.52	62.44	0.85	0.85	29.39	57.41	0.84	0.85
	PLUME 128k	33.25	59.78	0.83	0.83	23.91	52.93	0.80	0.81
	PLUME 256k	33.80	59.69	0.83	0.82	24.72	53.34	0.81	0.81
	PLUME 32k	32.60	58.97	0.82	0.82	24.11	52.80	0.80	0.80
pt-gl	NLLB 3.3B	31.12	57.92	0.88	0.83	32.55	59.00	0.87	0.82
	PLUME 128k	28.83	56.91	0.87	0.82	28.27	56.48	0.85	0.81
	PLUME 256k	28.58	56.52	0.87	0.82	28.54	56.57	0.85	0.81
	PLUME 32k	28.64	56.61	0.87	0.82	28.01	56.32	0.85	0.81
pt-it	NLLB 3.3B	26.42	55.44	0.88	0.85	31.19	59.11	0.87	0.85
	TOWERBASE 7B	22.31	52.69	0.85	0.85	31.19	59.11	0.87	0.85
	PLUME 128k	24.06	53.75	0.86	0.84	26.97	56.30	0.85	0.83
	PLUME 256k	24.24	53.75	0.86	0.84	27.46	56.52	0.85	0.83
	PLUME 32k	23.67	53.46	0.85	0.83	27.60	56.49	0.85	0.83

Improving Fluency Of Neural Machine Translation Using Large Language Models

Jianfei He, Wenbo Pan, Jijia Yang, Sen Peng, Xiaohua Jia
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
{jianfeihe-2c, wenbo.pan, jijiayang2-c, senpeng.cs}@my.cityu.edu.hk
csjia@cityu.edu.hk

Abstract

Large language models (LLMs) demonstrate significant capabilities in many natural language processing tasks. However, their performance in machine translation is still behind that of the models specially trained for machine translation with an encoder-decoder architecture. This paper investigates how to improve neural machine translation (NMT) with LLMs. Our proposal is based on an empirical insight that NMT gets worse fluency than human translation. We propose to use LLMs to enhance the fluency of NMT's generation by integrating a language model at the target side. We use contrastive learning to constrain fluency so that it does not exceed the LLMs' fluency. Our experiments on three language pairs show that this method can improve the performance of NMT. Our empirical analysis further demonstrates that this method improves the fluency on the target side. Our experiments also show that some straightforward post-processing methods using LLMs, such as re-ranking and refinement, are not effective.

1 Introduction

Large Language Models (LLMs) such as GPT (Ouyang et al., 2022; Achiam et al., 2023) and LLama (Touvron et al., 2023; Dubey et al., 2024) have demonstrated significant capabilities in various domains, including language understanding and generation tasks (Chang et al., 2024). However, the evaluations (Hendy et al., 2023; Zhu et al., 2024) show that LLMs' performance in machine translation is still behind the models dedicated to the task. These dedicated models often use an encoder-decoder architecture and are trained with parallel corpora. This raises a question: Can LLMs still help improve neural machine translation (NMT)?

A key translation challenge is the balance between *adequacy* and *fluency*. According to Läubli et al. (2018), NMT is good at adequacy and weak at fluency compared to human translation. There are some *post-processing* methods to use LLMs on NMT's outputs to improve fluency. We can follow the *reranking* methods in NMT (Lee et al., 2021; Bhattacharyya et al., 2021; Fernandes et al., 2022). LLMs can be used to rerank the candidates that are output from NMT, and the one with the smallest perplexity, according to LLM's evaluation, is chosen as the final output. Alternatively, we apply the *self-refine* method in LLM (Pan et al., 2023; Li et al., 2024; Han et al., 2024) to NMT's outputs. The translations from NMT are included in the prompt and an LLM is explicitly asked to refine their fluency. These two methods are used as baselines in our experiments. Results show that they cannot consistently improve the performance of NMT.

We propose to improve the fluency of NMT's translation by integrating the language capability of LLMs during training the NMT model. We use a two-pass strategy in the decoder. The first pass is a normal one using parallel sentences. The second pass only uses the target sentences in the training data. The objective is to train a target language model while training the translation model. This is realized by assigning *all ones* to the context vectors from the encoder for the second pass. Furthermore, we use an LLM to infer the training set and get their negative log-likelihoods. These data are used with *contrastive learning* to constraint the fluency of the target language model not to exceed the LLM's.

We conduct experiments on three language pairs: German-English (De–En), Russian-English (Ru–En), and French-English (Fr–En). The results show that our method effectively improves the performance of NMT. Our empirical analysis further demonstrates that our method improves fluency on the target side, and contrastive learning with

Proceedings of Machine Translation Summit XX Volume 1, pages 54–64, June 23-27, 2025

knowledge from the LLM plays an important role in achieving gains.

2 Related Work

2.1 LLMs for Translation

There is a line of research to use prompt engineering and few shot learning for LLM to translate (Zhang et al., 2023a; Gao et al., 2023). Evaluations (Hendy et al., 2023; Zhu et al., 2024) show that LLMs' performance in machine translation is still behind the NMT models dedicated to this task.

Zhang et al. (2023b), Alves et al. (2024) and Xu et al. (2024) also explore finetuning LLMs with parallel corpora to get better performance. Since LLMs have a much larger number of parameters than typical NMT, finetuning these models with a dedicated parallel corpus is not a convincing method. Such a method also does not follow the paradigm of LLMs, which aims to be general for many tasks instead of one specific downstream task.

Reranking is well investigated in the context of NMT (Lee et al., 2021; Bhattacharyya et al., 2021; Fernandes et al., 2022). The reranker is either a reference-free evaluation method such as COMET (Fernandes et al., 2022) or a dedicated trained score model in Lee et al. (2021). To the best of our knowledge, there is no research using LLMs to reranking NMT. We implement this method as one baseline in our experiments.

Using LLM to refine its own output has been investigated and is effective for some NLP tasks other than translation (Pan et al., 2023; Li et al., 2024; Han et al., 2024). Bogoychev and Chen (2023) use LLM to refine NMT's results. Their research focuses on a specific use case: terminology-aware translation.

2.2 Contrastive Learning (CL) in NLP

Contrastive Learning is applied to NMT by Yang et al. (2019) and Pan et al. (2021). However, they address specific issues. Yang et al. (2019) aim to reduce the word omission errors and Pan et al. (2021) use CL to improve the many-to-many multilingual NMT. We aim to improve the fluency of NMT, which is a more general objective.

Besides NMT, CL has applications in other NLP tasks. Sun and Li (2021) and Liu et al. (2022) apply CL for text summarization. Sun and Li (2021) use a pair-wise preference. The gold references are positive samples, and low-quality predictions are negative ones. Liu et al. (2022) use a list-wise preference. A group of ranked predictions are used in CL. These two methods work at the sequence level, while ours works at the token level.

Su et al. (2022) aim to mitigate the anisotropic distribution of token representations. They use CL to calibrate the representation space for tokens in the model.

3 Methodology

3.1 Adequacy and Fluency

Our proposal is based on the insight that NMT gets worse fluency than human translation.

There are two goals for machine translation: fluency and adequacy (Läubli et al., 2018; Kong et al., 2019; Miao et al., 2021; Sulem et al., 2020). Fluency measures whether a translation is fluent in terms of the target language. Adequacy measures whether the translation conveys the correct meaning in the source sentence, even if the translation is not fully fluent viewing from the target language.

While adequacy often requires human evaluation, fluency can be easily evaluated using the *perplexity* (denoted as *ppl*) with a language model at the target side. The relationship between perplexity and NLL (Jurafsky and Martin, 2020) is :

$$\begin{aligned} NLL &= -\sum_{i=1}^{n} log\ p(y_i|y_{<i}), \\ ppl &= e^{NLL} \end{aligned} \tag{1}$$

where y_i is the i^{th} target token and n is the total length of the target sentence.

According to Läubli et al. (2018), NMT is good at adequacy and weak at fluency compared to human evaluation. Their main result is illustrated in Figure 1.

3.2 Two-Pass Decoder

We use a two-pass procedure in the decoder in training. Each pass is related to a component in the loss function.

The first pass is through a standard decoder and gets the usual loss value of maximum likelihood estimation (MLE), which is the negative log-likelihood (NLL) with label smoothing (Edunov et al., 2018):

$$\begin{aligned} \mathcal{L}_{MLE} = &-\sum_{i=1}^{n} log\ p(y_i|X, y_{<i}) \\ &-D_{KL}(f \parallel p(y_i|X, y_{<i})), \end{aligned} \tag{2}$$

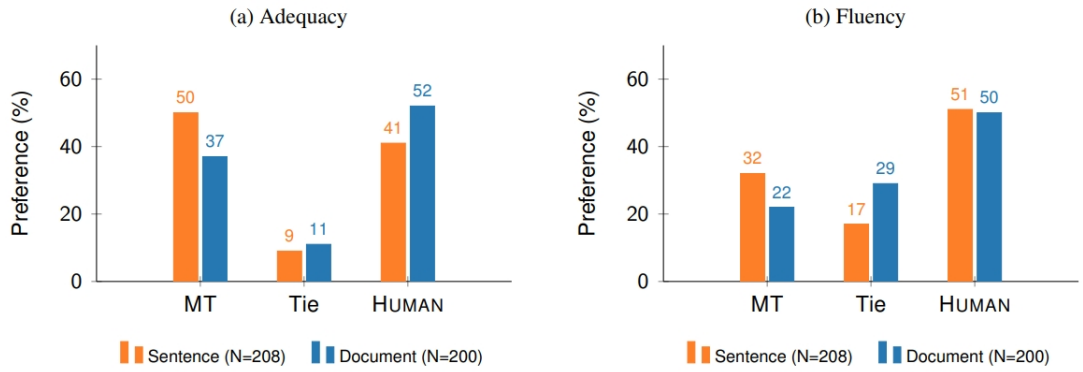


Figure 1: There is no statistically significant difference between HUMAN (human translation) and MT in terms of adequacy when evaluating sentences. However, raters show a significant preference for HUMAN in terms of fluency. From Läubli et al. (2018)

.

where X and y_i denote the source sentence and the ground truth token for step i, respectively, and f is the uniform distribution over the vocabulary. When the size of the vocabulary is V, $f = \frac{1}{V}$.

The objective of the second pass is to train the decoder to learn a target language model by *turning off* the context attention. It is realized by assigning *all ones* to the values of context vectors from the encoder. In this way, the cross-attention reduces to the query from the decoder side:

$$\begin{aligned} &Attention(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = softmax(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d_k}})\mathcal{V} \\ &= softmax(\frac{\mathcal{Q}}{\sqrt{d_k}}),\ when\ \mathcal{K}, \mathcal{V}\ are\ all\ ones. \end{aligned} \tag{3}$$

Correspondingly, this second pass gets the second loss component:

$$\mathcal{L}_{fluency} = -\sum_{i=1}^{n} log\ p(y_i|y_{<i}) \tag{4}$$

In this two-pass procedure, the same network architecture is used, and all parameters are shared.

This is a potential conflict between the $\mathcal{L}_{fluency}$ in Equation 4 and $\mathcal{L}_{MLE}$ in Equation 2. When the model is trained using the loss component in Equation 4, $log\ p(y_i|y_{<i})$ is maximized. This may conflict with the translation objective in Equation 2 which maximizes $log\ p(y_i|y_{<i}, X)$. We use *contrastive learning* to mitigate this conflict.

3.3 Contrastive Fluency Enhancement (CFE)

Contrastive Learning (CL) has a key component: a *max* function. It is defined as:

$$\max\{0, \rho + S_n - S_p\}, \tag{5}$$

where S_n and S_p are scores for negative and positive samples, respectively. ρ is a hyperparameter, the margin between the scores between negative and positive samples.

This function outputs a positive loss when the score of the negative sample is larger than one margin plus the score of the positive sample. The objective is to constrain the score of the negative sample so that it is at least one margin lower than the score of the positive sample.

We use the negative log-likelihood (NLL) of target tokens as the scores. The values from the training models are negative samples, while those from LLMs are positive samples. This method is denoted as Contrastive Fluency Enhancement (CFE) and the corresponding loss component is:

$$\begin{aligned} &\mathcal{L}_{CFE} = \\ &\max\{0, \rho - \sum_{i=1}^{n} log\ p(y_i|y_{<i}) + \sum_{i=1}^{n} log\ p_{llm}(y_i|y_{<i})\}, \end{aligned} \tag{6}$$

where p_{llm} is the probability in LLM.

The final loss function is:

$$\mathcal{L} = \mathcal{L}_{MLE} + \mathcal{L}_{CFE} \tag{7}$$

To conduct the ablation study, we implemented a variant without CL. Its loss function is:

$$\mathcal{L}_{MLE} + \mathcal{L}_{fluency} = \mathcal{L}_{MLE} - \sum_{i=1}^{n} log\ p(y_i|y_{<i}) \tag{8}$$

4 Experiments

4.1 Datasets

We use the negative log-likelihood (NLL) from an LLM as the positive samples during training. The

major data used to train an LLM are usually in English. Therefore, we use English as the target language in our experiments.

We use the corpora from WMT[1] as our datasets.

We use Europarl v7, News-commentary-v12, and Common Crawl for training in De–En. The training data have totally 4.6 million sentences. We use Newstest2014 for validation, and Newstest2021 for testing in De–En. For Ru–En, ParaCrawl v9, News-commentary-v10, and Common Crawl are used for training. These training data have totally 13.1 million sentences. Newstest2014 is used for validation, Newstest2021 is used for testing in Ru–En. For Fr–En, Europarl v7, News-commentary-v10, and Common Crawl are used for training. These training data have totally 5.4 million sentences. Newstest2013 is used for validation, and Newstest2015 is used for testing in Fr–En.

We need to use an LLM to infer each target sentence in the training set to get its negative log-likelihood. Therefore, we limit the size of the training set by filtering the original datasets. We randomly select 350 million sentences from the original training dataset for each language pair. We use the condition below to further choose data with high quality:

- Both source and target sentences have lengths within the range of 5 to 300.
- The disparity between the source and target sentence length does not exceed five times.

The number of sentence pairs for each language pair is as follows: De–En 2.6 million, Ru–En 2.9 million, Fr–En 2.7 million.

4.2 Systems

We compare our method with the vanilla Transformer model, three typical token-level methods improving NMT, and two methods introduced in Section 2 for comparison. Our method is not compared with sequence-level methods such as *MIXER* (Ranzato et al., 2016) and *MRT* (Shen et al., 2016). These sequence-level methods use online samples and are more than ten times slower than the token-level methods (Edunov et al., 2018).

- *TX* is the vanilla Transformer.
- *SS* (Mihaylova and Martins, 2019) is a scheduled sampling method with a Transformer that uses two-pass decoding. The Inverse Sigmoid Decay is used for scheduling in our experiments. It performs best among the scheduling algorithms according to Liu et al. (2021).
- *CASS* (Liu et al., 2021) is Confidence-Aware Scheduled Sampling. It enhances the normal scheduled sampling by sampling different tokens according to the model's probability of ground truth tokens.
- *TFN* (Goodman et al., 2020) uses two stacking decoders. The loss values are computed on each decoder and the results are combined to form the final loss value. We use the hyperparameters according to their recommendation in the paper. The second decoder's weight is set to 0.4, and both decoders share the same set of parameters.
- *Refine* includes the translations from NMT in the prompt and explicitly asks LLM to refine the fluency.
- *ReRank* uses LLMs to rerank the output candidates from NMT and choose the one with the smallest perplexity in LLM.

We implement our proposal, Contrastive Fluency Enhancement (CFE), as described in Section 3.

Since *ReRank* is a post-processing method, we can apply ReRank to the output of CFE. This variant is denoted as *CFE+ReRank*.

4.3 Implementation Details

We use Llama2-13B-chat-hf [2] as the LLM for experiments. Its negative log-likelihood of each token in the target sentences in the training data is used as described in Section3.3. For the method *Refine*, this model is also used to generate refined translations. In inference, we use top-p (0.9) sampling, and the sampling temperature is set to 0.9.

Our implementation of NMT is based on the Fairseq toolkit (Ott et al., 2019) using a typical configuration [3] similar to the original Transformer (Vaswani et al., 2017). The Transformer Base model with about 60 million parameters is used. Since we use the token-level negative log-likelihood from Llama2-13B-chat-hf, we need to

[1] http://www.statmt.org

[2] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

[3] https://github.com/facebookresearch/fairseq/tree/main/examples/scaling_nmt

	De–En			Ru–En			Fr–En		
Metrics	BLEU	Meteor	Comet	BLEU	Meteor	Comet	BLEU	Meteor	Comet
Baselines									
Transformer	26.19	49.18	75.45	28.76	49.98	75.28	34.41	51.17	76.51
SS	26.43	49.20	75.40	28.71	49.82	74.99	34.55	51.12	76.29
CASS	26.27	**49.54**	75.56	28.96	50.14	75.30	35.14	51.72	76.67
TFN	26.31	**49.54**	75.44	28.99	50.23	75.30	34.32	51.13	76.67
Refine	26.19	49.18	75.45	28.76	49.98	75.28	34.41	51.17	76.51
ReRank	26.42	**49.90**	75.76	28.99	**51.05**	75.93	*33.09*	*51.08*	*76.06*
Δ (-TX)	0.23	0.72	0.31	0.23	1.07	0.65	-1.32	-0.09	-0.45
Our Proposal									
CFE	**26.65**	**49.45**	**75.91**	**29.67**	**50.82**	**76.51**	**35.50**	**51.88**	**76.86**
Δ (-TX)	0.46	0.27	0.46	0.91	0.84	1.23	1.09	0.71	0.35
CFE+ReRank	**27.06**	**50.18**	**76.03**	**29.72**	**51.73**	**76.87**	*33.76*	51.74	*76.12*
Δ (-TX)	0.87	1.00	0.58	0.96	1.75	1.59	-0.65	0.57	-0.39

Table 1: Performance of different methods. The scores of CFE and those better than CFE are highlighted in **Bold**, while the scores that are worse than the vanilla Transformer (denoted as **TX**) are shown in *Italic*. Δ denotes the gain compared to TX.

use the same tokenizer for NMT and Llama2-13B-chat-hf so that one sentence has the same subwords in two systems. We use the tokenizer of Llama2-13B-chat-hf for subwords. The vocabulary size is equal to 32,000, which is shared for the source and target sentences. Both the dropout rate and the label smoothing are set to 0.1. We use beam search for decoding with a beam size of six, and the factor for length penalty is 0.6. The number of candidates used for *ReRank* is the same as the beam size.

In our preliminary experiments for *Refine*, we found that the outputs from LLMs may contain some explanation words. This result makes it difficult to extract the refined sentence for evaluation. Therefore, the prompt used for *Refine* in our evaluation requires that the LLM do not give any explanation. The prompt is shown below:

"initial translation"

If there are minor mistakes in the above sentence, please correct them and make this sentence more fluent. If there is no mistake, keep it intact. Only output the result. No explanation.

Our method, its variant for ablation study, and token-level baseline methods (SS, CASS, TFN) use a common pre-trained NMT model for finetuning. This pre-trained model is trained for a minimum of 20 epochs on the filtered data set described in Section 4.1, stopping if the validation loss does not decrease for 20 consecutive epochs. For finetuning, we adopt the same early-stop policy as Choshen et al. (2019), where the process is terminated if the validation loss does not decrease for ten consecutive epochs. The margin ρ in the loss function of CFE is set to 0.1.

All GPUs used for training are Nvidia GF1080Ti.

4.4 Evaluation and Results

Three metrics are used to evaluate the performance of the methods using: BLEU, Meteor, and Comet. We use SacreBLEU[4] (Post, 2018)[5] for BLEU. For Meteor[6], we use its version 1.5. For Comet, we use its *wmt22-comet-da* model[7].

Table 1 illustrates the performance of methods for De–En, Ru–En, and Fr–En.

The vanilla Transformer model is a strong baseline. Our method CFE outperforms it in all three metrics for all language pairs. CPE generally achieves the best performance compared to other baselines except for a few cases in Meteor.

Refine gets the same performance as the vanilla Transformer. We find that LLM almost always regards the translation from NMT as fluent enough and does not provide improved translations. The number of *intact* sentences are illustrated in Table 4.

ReRank gets better performance than the vanilla Transformer for De–En and Ru–En, but much

[4] https://github.com/mjpost/sacreBLEU
[5] case.mixed+numrefs.1+smooth.exp+tok.13a+version.2.3.1
[6] http://www.cs.cmu.edu/~alavie/METEOR/
[7] https://github.com/Unbabel/COMET

worse for Fr–En. Table 2 and 3 illustrate that ReRank always gets much lower perplexity than the vanilla Transformer. The inconsistency between low perplexity and good translation reflects the complexity of machine translation and the importance of the balance between adequacy and fluency.

CPE+ReRank gets gains in De–En and Ru–En. However it has worse performance than CPE in Fr–En. This result is consistent with the bad performance of ReRank alone in Fr–En.

	Model	De–En	Ru–En	Fr–En
ppl	TX	217.9	128.5	242.3
	ReRank	73.0	62.4	94.9
	CFE	117.6	131.5	223.0
	CFE+ReRank	72.1	66.1	87.8
NLL	TX	4.131	3.923	4.406
	ReRank	3.823	3.631	4.019
	CFE	4.108	3.895	4.404
	CFE+ReRank	3.798	3.602	3.999

Table 2: Fluency measured with average perplexity (ppl) and negative log-likelihood (NLL).

	De–En	Ru–En	Fr–En
Better	855	835	1301
Equal	145	165	195
Worse	0	0	0

(a) ReRank, compared to Transformer

	De–En	Ru–En	Fr–En
Better	477	486	578
Equal	95	93	346
Worse	428	421	572

(b) CFE, compared to Transformer

	De–En	Ru–En	Fr–En
Better	775	767	1174
Equal	37	38	113
Worse	188	195	209

(c) CFE+ReRank, compared to Transformer

Table 3: Investigate the fluency compared to Transformer at sentence-level using negative log-likelihood.

5 Analysis

5.1 Loss Components in CFE

Figure 2 shows the components in the loss function of CFE for De–En during training. Both the loss component $\mathcal{L}_{fluency}$(Figure 2a) and the total loss (Figure 2b) steadily decrease. These figures demonstrate the effectiveness of the CFE loss function presented in Section 3.

The results on other language pairs get to the same conclusion as illustrated in Figure 3 and 4.

5.2 Fluency

The fluency usually is measured with *perplexity*, denoted as *ppl*. We use Llama2-13B-chat-hf to get the NLL of each translation, which is averaged based on the number of tokens in the generated sentence. These NLLs are used to calculate that sentence's perplexity according to Equation 1.

Table 2 illustrates each test set's average perplexity and NLL. ReRank outputs the one with the lowest NLL in the candidates. Therefore, it consistently gets much lower perplexity and NLL compared with the vanilla Transformer, even for Fr–En that ReRank gets much worse performance as shown in Table 1.

Our method CFE consistently gets lower NLL for all language pairs than the vanilla Transformer. CFE generally gets a lower average perplexity, with the only exception being Ru–En. Compared to ReRank, CFE gets larger perplexity and NLL. This result reflects that CFE gets a better balance between fluency and adequacy.

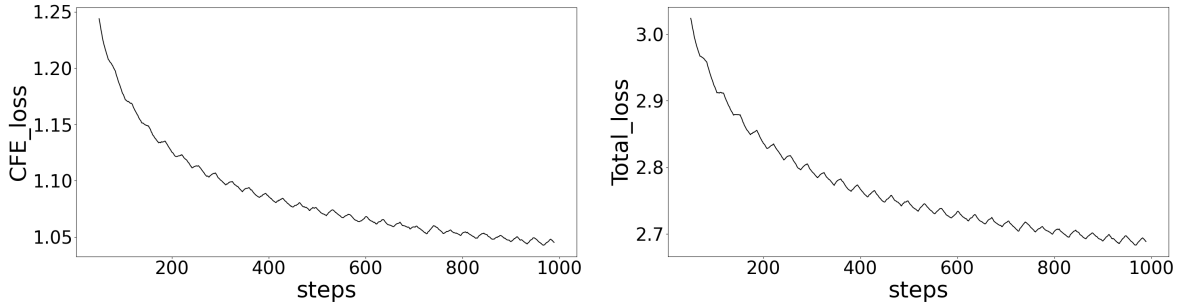
We also compare the NLL of the vanilla Transformer and other methods for each translation and count the number of cases that other methods have lower (*better*), equal, or greater (*worse*) NLL than the vanilla transformer. When the absolute value of the difference in comparison is less than 0.001, two NLL values are counted as *equal*. The results illustrated in Table 3 show that our method effectively improves the fluency of NMT.

5.3 Refine With LLM

Table 4 illustrates that most sentences are kept intact when the LLM is asked to improve fluency. There are a few sentences in which no translations are identified in the feedback from Llama2-13B-chat-hf. When these *empty* feedback are identified, the original translations are reasonably used before evaluation in our implementation. This analysis explains why *Refine* gets the exactly same performance as the vanilla Transformer.

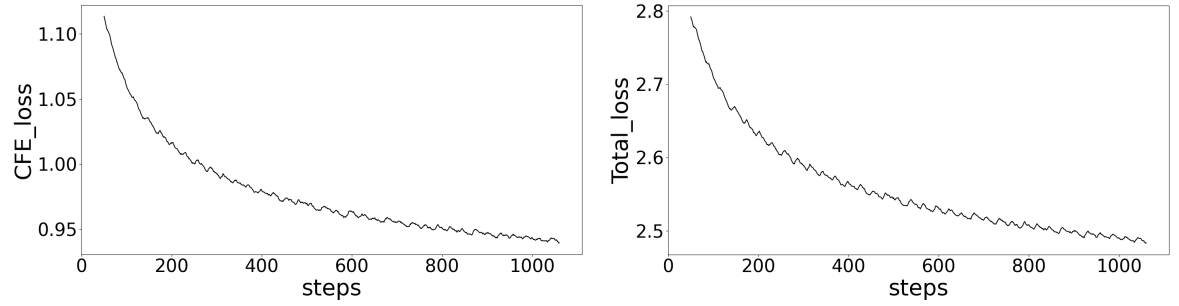
5.4 Ablation Study

Table 5 shows the performance of CFE with and without Contrastive Learning (CL). The variant without CL implements the loss function in Equation 8. It maximizes the target language model



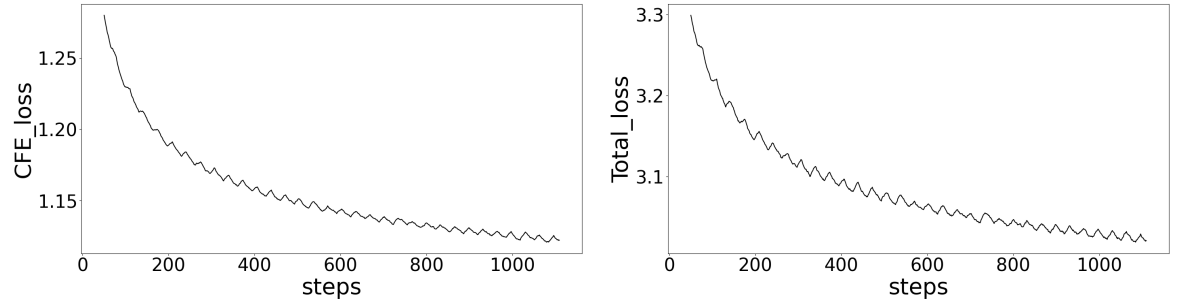
(a) Loss component of CFE (b) Total loss

Figure 2: Investigate the components in the loss function for De–En



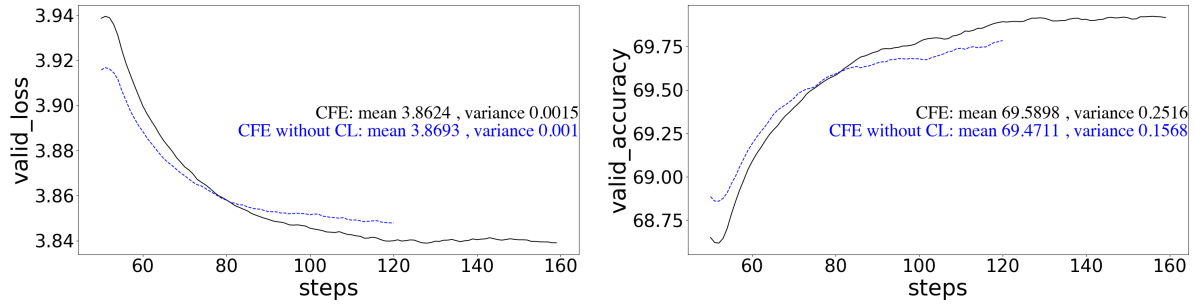
(a) Loss component of CFE (b) Total loss

Figure 3: Investigate the components in the loss function for Ru–En



(a) Loss component of CFE (b) Total loss

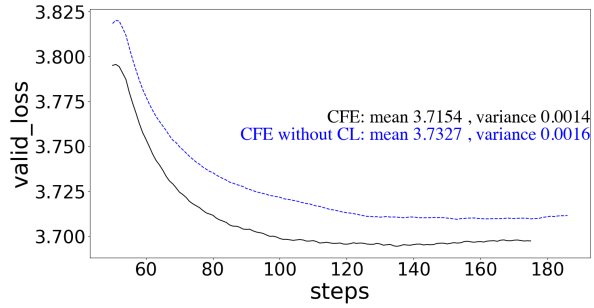
Figure 4: Investigate the components in the loss function for Fr–En



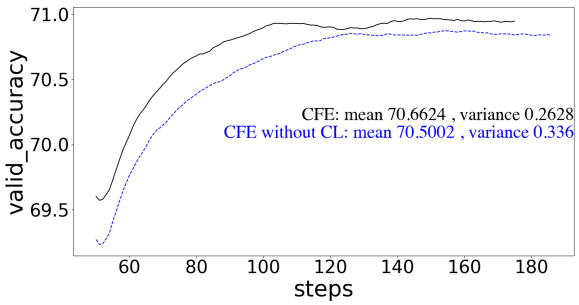
(a) Loss on validation set (b) Accuracy on validation set

Figure 5: Investigate the performance on the validation sets during training for CFE and its variant without contrastive learning for De–En.

and does not make use of LLM's knowledge as a ceiling. While CFE without CL also outperforms the vanilla Transformer model and demonstrates its efficacy in improving NMT, its gains are generally

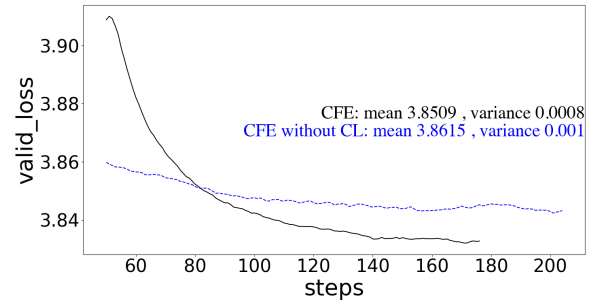


(a) Loss on validation set

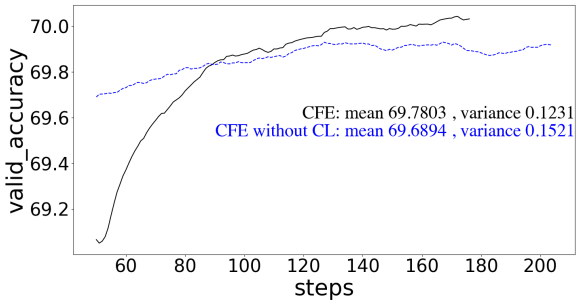


(b) Accuracy on validation set

Figure 6: Investigate the performance on the validation sets during training for CFE and its variant without contrastive learning for Ru–En.



(a) Loss on validation set



(b) Accuracy on validation set

Figure 7: Investigate the performance on the validation sets during training for CFE and its variant without contrastive learning for Fr–En.

	De–En	**Ru–En**	**Fr–En**
Total	1000	1000	1496
Intact	995	996	1480
Empty	5	4	16

Table 4: *Refine* with the LLM does not improve NMT.

Model	**De–En**	**Ru–En**	**Fr–En**
TX	26.19	28.76	34.41
CFE	26.65	29.67	35.50
Δ **(-TX)**	0.46	0.91	1.09
w/o-CL	26.58	29.01	34.66
Δ **(-TX)**	0.39	0.25	0.25

Table 5: Ablation test by removing Contrastive Learning from CFE, denoted as *w/o-CL*.

lower than CFE.

Figure 5 shows the performance on the validation sets during training for CFE (black and solid) and its variant (blue and dashed) without CL in De–En. It shows that the variant consistently gets higher loss and lower accuracy during training. Figure 6 and 7 illustrate the performance on the validation set for the other language pairs, which are consistent with the conclusion of De–En.

These ablation tests demonstrate the importance of Contrastive Learning in CFE.

5.5 Significance Tests

Table 6 shows the results of significance tests for *ReRank*, *CFE* and *CFE+ReRank* (denoted as *CFE+RR+ST*). We report mean and standard error over five training runs with seeds 1–5. For *ReRank*, these seeds are applied to pretrained models. These results are generally consistent with Table 1.

Model	**BLEU**	**Meteor**	**Comet**
TX	26.19	49.18	75.45
ReRank-ST	26.37 ±.11	49.88 ±.09	75.70 ±.06
Δ **(-TX)**	0.18	0.70	0.25
CFE-ST	26.65±.09	49.37 ±.06	75.87 ±.07
Δ **(-TX)**	0.46	0.19	0.42
CFE+RR-ST	26.70±.11	49.84±.11	75.85 ±.09
Δ **(-TX)**	0.51	0.66	0.40

Table 6: Significance tests on De–En.

6 Conclusion

This paper investigates how to improve neural machine translation (NMT) with Large language models (LLMs). Our experiments show that post-processing methods like re-ranking and self-refining are not effective. Based on the insight that NMT is good at adequacy and weak at fluency, we propose to use LLMs to enhance the fluency of NMT's generation by integrating a language model at the target side and using Contrastive learning to constraint the probabilities to a ceiling, the LLM's fluency. Our experiments on three language pairs (De–En, Ru–En, and Fr–En) show that this method effectively improves the performance of NMT. The empirical analysis further demonstrates that this method improves the fluency at the target side and Contrastive Learning with knowledge from the LLM plays an important role in achieving the gains.

7 Sustainability Statement

We trained and finetuned the model with the early-stop strategy as described in Section 4.3. Pretraining and finetuning the model typically took nearly 140 and 100 GPU-hours using Nvidia GF1080Ti. The estimated energy cost for each model is illustrated in Table 7, according to the calculation using Green-Algorithms[8].

	GPU-Hour	CO_2(kg)	Engergy(kWh)
Pretrain	140	31.88	59.32
Finetune	110	25.05	46.61

Table 7: Estimated energy cost for each model.

[8] http://calculator.green-algorithms.org

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Duarte M Alves, José Pombal, Nuno Miguel Guerreiro, Pedro Henrique Martins, João Alves, M Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *CoRR*.

Sumanta Bhattacharyya, Amirmohammad Rooshenas, Subhajit Naskar, Simeng Sun, Mohit Iyyer, and Andrew McCallum. 2021. Energy-based reranking: Improving neural machine translation using energy-based 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)*. 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.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Leshem Choshen, Lior Fox, Zohar Aizenbud, and Omri Abend. 2019. On the weaknesses of reinforcement learning for neural machine translation. In *International Conference on Learning Representations*.

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*.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. 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 355–364.

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.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. How to design translation prompts for chatgpt: An empirical study. *Preprint*, arXiv:2304.02182.

Sebastian Goodman, Nan Ding, and Radu Soricut. 2020. Teaforn: Teacher-forcing with n-grams. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8704–8717.

Haixia Han, Jiaqing Liang, Jie Shi, Qianyu He, and Yanghua Xiao. 2024. Small language model can self-correct. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18162–18170.

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*.

Daniel Jurafsky and James H Martin. 2020. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

Xiang Kong, Zhaopeng Tu, Shuming Shi, Eduard Hovy, and Tong Zhang. 2019. Neural machine translation with adequacy-oriented learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6618–6625.

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.

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, Online. Association for Computational Linguistics.

Loka Li, Guangyi Chen, Yusheng Su, Zhenhao Chen, Yixuan Zhang, Eric Xing, and Kun Zhang. 2024. Confidence matters: Revisiting intrinsic self-correction capabilities of large language models. *CoRR*.

Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Confidence-aware scheduled sampling for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2327–2337.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. Brio: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903.

Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. 2021. Prevent the language model from being overconfident in 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 3456–3468.

Tsvetomila Mihaylova and André FT Martins. 2019. Scheduled sampling for transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 351–356.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual 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 244–258.

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.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016*.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. *Advances in Neural Information Processing Systems*, 35:21548–21561.

Elior Sulem, Omri Abend, and Ari Rappoport. 2020. Semantic structural decomposition for neural machine translation. In *Proceedings of the ninth joint conference on lexical and computational semantics*, pages 50–57.

Shichao Sun and Wenjie Li. 2021. Alleviating exposure bias via contrastive learning for abstractive text summarization. *arXiv preprint arXiv:2108.11846*.

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*.

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.

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*.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023b. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with qlora. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481.

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.

Optimizing the Training Schedule of Multilingual NMT using Reinforcement Learning

Alexis Allemann[1], Àlex R. Atrio[2,*], Andrei Popescu-Belis[1,3]

[1]HEIG-VD / HES-SO, 1401 Yverdon-les-Bains, Switzerland
[2]Pi School, Rome, Italy
[3]EPFL, 1015 Lausanne, Switzerland
Correspondence: andrei.popescu-belis@heig-vd.ch

Abstract

Multilingual NMT is a viable solution for translating low-resource languages (LRLs) when data from high-resource languages (HRLs) from the same language family is available. However, the training schedule, i.e. the order of presentation of languages, has an impact on the quality of such systems. Here, in a many-to-one translation setting, we propose to apply two algorithms that use reinforcement learning to optimize the training schedule of NMT: (1) Teacher-Student Curriculum Learning and (2) Deep Q Network. The former uses an exponentially smoothed estimate of the returns of each action based on the loss on monolingual or multilingual development subsets, while the latter estimates rewards using an additional neural network trained from the history of actions selected in different states of the system, together with the rewards received. On a 8-to-1 translation dataset with LRLs and HRLs, our second method improves BLEU and COMET scores with respect to both random selection of monolingual batches and shuffled multilingual batches, by adjusting the number of presentations of LRL vs. HRL batches.

1 Introduction

Multilingual neural machine translation (NMT)is particularly effective to enable the translation of low-resource languages (LRLs) when they are accompanied, in the training data, by related high-resource languages (HRLs) (Gu et al., 2018; Neubig and Hu, 2018). Including HRLs in the training data reduces the chance of overfitting to the LRLs and improves translation quality.

Many-to-one NMT systems can be trained either with monolingual or with multilingual batches. Monolingual batches include a single language on the source side, while multilingual batches have their source side sampled from several source languages. Using multilingual batches helps avoiding catastrophic forgetting (Jean et al., 2019), but the mixture of languages in each batch may be ineffective at early stages of training. Here, we focus on monolingual batches, as they enable us to define the training schedule of a NMT system as the order of presentation of languages, but we also compare our results to those obtained with multilingual batches.

We propose to use reinforcement learning (RL) to optimize the training schedule of many-to-one NMT systems, i.e. to improve the training process and the resulting system compared to a fixed sampling strategy. We enable our systems to select the source language of the batch at each training step, based on a learned estimate of the model's competence on each language in terms of loss on a development set. Unlike fixed strategies, such as training on the hardest language, we leverage RL to let the model find better strategies.

We make the following contributions:[1]

- We apply the Teacher-Student Curriculum Learning algorithm (Matiisen et al., 2017) to NMT by modeling the expected return as the smoothed loss of the NMT system over a development set.
- Based on the Deep Q Network algorithm (DQN) (Mnih et al., 2013), we design a RL-based model in which the expected rewards are generated by an auxiliary neural network trained in parallel with the NMT system.
- We perform experiments on a dataset with four language families on the source side, with one HRL and one LRL for each family (Neubig and Hu, 2018); the target language is English.
- DQN outperforms in terms of BLEU and

*Work performed while at HEIG-VD / HES-SO.

[1]Source code is made available at https://github.com/alexis-allemann/OpenNMT-py/tree/curriculum_learning.

Proceedings of Machine Translation Summit XX Volume 1, pages 65–80, June 23-27, 2025

COMET scores previous training schedules used for multilingual NMT: monolingual minibatches sampled equally, or in proportion of each language, or multilingual batches.

- Algorithms are robust to the setting of hyperparameters, and increase the proportions of LRLs in the training schedule from less than 1% to at least 4% while decreasing those of HRLs.

2 Related Work

Curriculum learning. In many applications of machine learning, the order of presentation of items from the training set may influence the outcome of the training, i.e. the quality of the final model, or the training speed. For instance, presenting items by increasing levels of difficulty is often beneficial, an approach known as *curriculum learning* (Wang et al., 2021). The difficulty can be measured directly on the data, or it can be inferred from the observed competence of the model during training, an approach known as *self-paced learning* (Kumar et al., 2010; Jiang et al., 2015). The competence of a model can be estimated intrinsically, e.g. from its loss values on a subset of the data, or extrinsically, by using a teacher model that observes the behavior of the target model, called 'student' (Matiisen et al., 2017). Competence can be used by the teacher model to adjust the training schedule of the student model. In the case of systems that can perform several tasks, the training schedule consists of the selection of tasks and related data.

When the teacher model is in charge of the training schedule of the student, it may use reinforcement learning (RL), with the student model playing the role of the environment (Shen and Zhao, 2024). RL has proved particularly useful at training large language models to follow instructions (Ouyang et al., 2022), initially using PPO (Proximal Policy Optimization, Schulman et al., 2017) and then other algorithms (Rafailov et al., 2023; Ethayarajh et al., 2024), but these methods are not designed to optimize training schedules. While it is possible to use curriculum learning to train RL-based models (Narvekar et al., 2020), e.g. by presenting them with increasingly difficult problems, we focus here on the use of RL to train a teacher model, in the field of multilingual NMT.

Training schedules for NMT. Optimizing the training schedule of an NMT system depends in particular on its architecture. For a system with a single input language and domain, the training sentences can be presented by order of estimated difficulty, or by order of translation reliability or noisiness. When multiple domains must be considered, additional decisions must be made about which domain to use first, or how to mix them based on sizes of available data. Similar decisions must be made if there are multiple input languages, as in our case, or if one must train a multi-task system including NMT along with other tasks such as language modeling. We briefly review here previous work along these lines.

Static scheduling in multilingual NMT. Neubig and Hu (2018) study the upsampling of the LRL data when building minibatches, and observe that keeping the original proportions of HRL and LRL performs marginally better. However, Johnson et al. (2017) and Aharoni et al. (2019) sample each batch uniformly from a concatenation of all language pairs. Arivazhagan et al. (2019) compare simple concatenation with uniform balancing, and observe better results for LRLs when using temperature-based upsampling, which was favored afterwards (Conneau et al., 2020; Tang et al., 2021).

The translation capabilities of large language models (LLMs) have also been explored: Zhu et al. (2024) compares several recent LLMs and shows that they can achieve state-of-the-art results when translating HRLs, but highlights their limitations in translating LRLs compared to NMT models. One of the leading open-weights LLMs for MT, Tower Instruct (Alves et al., 2024), is fine-tuned on a large set of translation-related tasks in 10 HRLs, with no particular scheduling of the fine-tuning data, and no reinforcement learning.

Curriculum learning in monolingual NMT. Self-pacing has been used in NMT at the sample level, for instance by estimating learning confidence as the variance across dropout runs, with better performance and faster convergence compared to human-designed schedules (Wan et al., 2020). Similarly, Liu et al. (2020) design a self-paced curriculum based on the norm of a token's embedding. Zhang et al. (2018) adopt a probabilistic view of curriculum learning and improve the convergence time of a DE-EN NMT system at no loss in translation quality, but no gain either; moreover, they note a high sensitivity to hyperparameter settings. Platanios et al. (2019) propose a scheduling criterion combining the difficulty of samples and the competence of the NMT model,

the latter estimated as a linear or square root function of the number of steps. This reduces training time by up to 70% and improves BLEU scores by 1–2 points on three different language pairs. Wang et al. (2018) extend domain-specific data selection methods to denoise NMT training, which significantly improves NMT performance on noisy data. Wang et al. (2020a) introduce a method for multi-domain data selection in NMT, using instance-level domain-relevance features and an automated training curriculum to enhance performance across multiple domains.

Curriculum learning in multilingual NMT. Jean et al. (2019) compare adaptively upsampling a language depending on various criteria, observing best results on LRLs when dynamically changing the norm of the gradient. Wang et al. (2020b) adaptively balance the languages by learning their weights from the model's competence on a development set. Zhang et al. (2021) design a dynamic sampling strategy which measures per-language competence but also evaluates LRL competence through a related HRL's competence. Wu et al. (2021) also balance the data dynamically, but measure a model's uncertainty as the variance over several runs of Monte Carlo dropout. Estimates of competence using the evolution of the loss of the NMT system have been proposed by Zaremoodi and Haffari (2019), who use its absolute value, by Xu et al. (2020), who use its relative decrease, and by Atrio et al. (2024), who use Kullback-Leibler divergence between consecutive states of the weights of an entire Transformer network.

RL-based curriculum learning in NMT. In the field of machine translation, Kumar et al. (2019) propose a RL framework utilizing Q-Learning to automatically learn an optimal curriculum for heterogeneous data, matching state-of-the-art hand-designed curricula. Zhao et al. (2020) introduce a RL-based data selection framework using Deterministic Actor-Critic to improve pre-trained NMT models by re-selecting influential samples from the original training set. Kreutzer et al. (2021) use a multi-armed bandit to dynamically select training data, thus optimizing NMT model performance across different domains, data qualities, and language pairs without manual schedule design.

Other applications of RL to NMT. In machine translation, RL methods were employed by Edunov et al. (2018) to tackle the discrepancy between token-level likelihood optimization during training and corpus-level evaluations using metrics like BLEU, and to reduce exposure bias in autoregressive sequence generators (Ranzato et al., 2016; Wang and Sennrich, 2020; Wu et al., 2018b). Kiegeland and Kreutzer (2021) emphasize the importance of exploration strategies, reward scaling, and reward function design for improving translation quality, particularly with respect to domain adaptation. To enhance the effectiveness of RL in NMT, Yehudai et al. (2022) show the importance of reducing the size and dimensionality of the action space. Wang et al. (2024) introduce efficient sampling-based RL techniques for sequence generation models, with a strong focus, however, on instruction tuning of LLMs.

3 Two RL Algorithms for Optimizing Training Schedules

In the RL framework, an agent observes the state S_t of the environment at each time step t, selects an action A_t based on its policy π, executes the action, and receives a reward R_t from the environment. Using the observed states, actions, and rewards, the goal is to learn an optimal policy, i.e. one that maximizes the cumulative reward over time. Bandit problems are those where the agent selects actions without considering state transitions.

In this study, as we use for training only monolingual batches, the set of possible actions $\mathcal{A}$ is simply the set of source languages. The states $\mathcal{S}$ of the system are the values of the parameters of the neural network and of the optimizer. However, these are too numerous to be sensibly observed at each step. Drawing inspiration from Wu et al. (2018a), we compute the current state of the model as the vector of cross-entropy loss values obtained from the NMT system over a development batch of sentences.[2] We use the score X_t of the model at time step t to compute the reward R_t as the decrease of the loss of the NMT system between the last two time steps: $R_t = X_t - X_{t-1}$. The loss values are computed on a development minibatch of data selected from the current language, or, in some experiments, on a multilingual minibatch.

3.1 TSCL Algorithm for NMT

Our first proposal is an adaptation to NMT of the Teacher-Student Curriculum Learning (TSCL) al-

[2]Alternatively, MT-specific metrics such as BLEU or COMET could be used instead of the cross-entropy loss, but computing them is more costly, therefore we do not use them here.

gorithm, a bandit method introduced by Matiisen et al. (2017), who use it to add decimal numbers or to navigate Minecraft mazes. The gist of our adaptation of TSCL for multilingual NMT is represented in Figure 1, and the full algorithm is given in Appendix A.4.

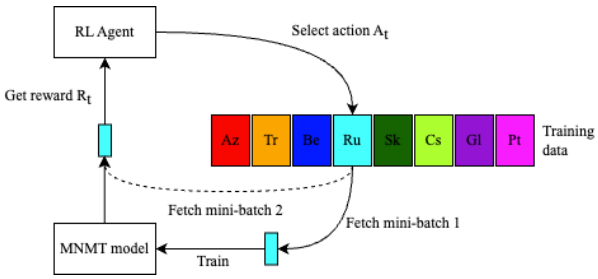


Figure 1: TSCL algorithm for NMT: relationship between the RL agent and the NMT system.

The action A_t of the system at time step t is the selection of a batch from a specific source language for training in the next step. The reward R_t of the action is the decrease of the negative cross-entropy loss of the model, X_t, computed on a batch of 8k tokens from the current source language, with respect to the value $X_{t'}$ computed at the latest previous time step with the same language. Formally, $R_t = X_t - X_{t'}$ where $t' = \max\{s : s < t \text{ and } A_s = A_t\}$. The expected return Q_t of the action is the exponentially weighted moving average of the rewards for the respective source language.

In some experiments, we start with a warm-up phase, a period during which all HRLs are randomly explored while the learning rate of the NMT model increases. Rewards of the RL agent only start to play a role after the warm-up phase, when the learning rate starts decreasing. In experiments without warm-up, each action is executed once at the beginning of the training, so that the model initiates training on the language that provides the highest reward from the start.

Additionally, to strike a balance between exploration and exploitation, we use an ϵ-greedy policy with a fixed value of ϵ. The action with the highest expected return is selected with probability $1 - \epsilon$, but with a small probability ϵ a random action is selected. In experiments with a warm-up period, this policy only starts after this period.

3.2 DQN Algorithm for NMT

The Deep Q Network (DQN) algorithm (Mnih et al., 2013) uses a neural network to approximate the Q-function that represents the expected reward of an action in a given state. The algorithm iteratively updates the parameters of this network to minimize the difference between the predicted Q-values and the desired Q-values obtained from the target system. Moreover, DQN enables experience replay by storing past experiences in a replay buffer and sampling them randomly during the training of the Q network, a feature that was shown to improve training.

Our application of DQN to multilingual NMT is illustrated in Figure 2, and the full algorithm is given in Appendix A.5. The RL agent is the Q network, a feed-forward neural network with *tanh* activation functions. Its input is the state of the NMT model: specifically, each value in the input layer represents the cross-entropy loss of the NMT model over a batch of 10 sentences from a specific language. Thus, an input vector of size 200 corresponds to a prototype batch of 2,000 sentences, with 250 sentences from each of the 8 source languages. The input layer is followed by two hidden layers of size 512 and by an output layer with 8 units, corresponding to the possible actions (selection of a source language for the next training step). The Q network is trained with the RMSProp optimizer[3] and the Huber Loss (Huber, 1964), a loss function that reduces the influence of extreme values, to mitigate the issue of outliers during training.

At each timestep t, the RL agent retains a new transition in its experience replay buffer. A transition consists of the previous state of the system S_{t-1}, the selected action A_{t-1}, the obtained reward R_t, and the current state of the system S_t. These transitions are used to train the Q network so that it predicts the action with the best estimated reward given the state of the NMT model.

We use an ϵ-greedy policy to balance between exploring actions and exploiting the Q network, like for TSCL. During the warm-up period, which is always applied to DQN, actions are randomly selected, but after it, actions are selected by the Q network with a probability of $1 - \epsilon$ or thy are randomly selected with a probability of ϵ. However, unlike TSCL, we follow Kumar et al. (2019) and start with $\epsilon = 1$ during warm-up, then gradually decrease this value at the end of warm-up to a minimum of 0.01 after 50k steps. This allows the network to randomly explore actions during

[3] A variant of stochastic gradient descent, proposed by G. Hinton, which adapts the learning rate for each parameter based on recent gradient averages (Ruder, 2016).

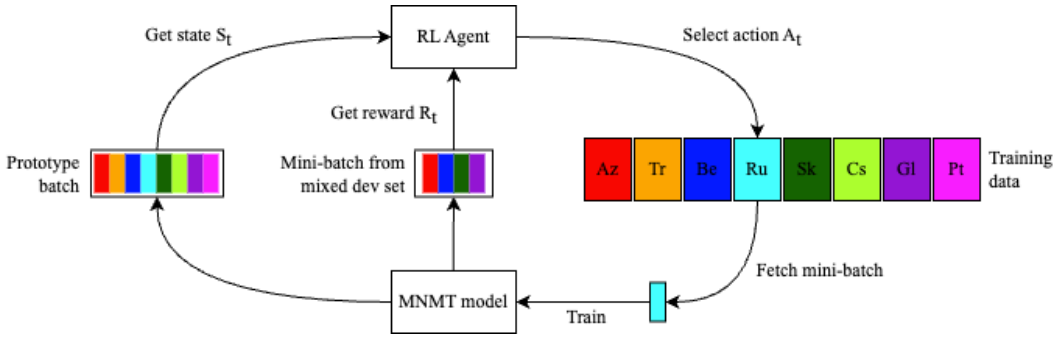


Figure 2: DQN algorithm for NMT: relationship between the RL agent (Q network) and the NMT model.

the warm-up period before exploiting the learned knowledge more and more. The schedule of ϵ is represented in Figure 5 of Appendix A.1.

4 Experimental Settings

Data. Experiments were conducted using a subset of the multilingual TED corpus collected by Qi et al. (2018), with four HRLs and four LRLs.[4] For comparability with prior research on multilingual NMT (Neubig and Hu, 2018; Wang et al., 2019; Zhang et al., 2021), we consider a 8-to-1 translation task with English as the target language. We are especially interested in the translation quality of the four LRLs of the dataset: Belarusian (BE), Azerbaijani (AZ), Galician (GL) and Slovak (SK), which are respectively paired with a HRL from the same family: Russian (RU), Turkish (TR), Portuguese (PT) and Czech (CS). Three language families are thus represented (Romance, Slavic and Turkic) but all scripts are Latin-based.

The numbers of sentences of the training and testing sets for each of the 8 languages are shown in Table 1. These numbers show that the distinction of LRLs vs. HRLs made in previous studies is to some extent arbitrary. Indeed, there are fewer PT sentences (considered nevertheless as a HRL with respect to GL) than SK sentences (considered as a LRL with respect to CS).

Preprocessing. The original data is already tokenized into words. We use Byte Pair Encoding (BPE) for subword extraction and vocabulary construction (Sennrich et al., 2016).[5] A vocabulary of 32k subwords is generated over a multilingual corpus obtained by combining 10k random lines from the training data of each language, with upsampling for AZ and BE which have fewer than 10k lines. For source language identification by the NMT model, each sentence is prefixed with a language tag.

Language	train	dev	test
AZ	5.9k (0.95%)	671	903
BE	4.5k (0.72%)	248	664
GL	10.0k (1.60%)	682	1.0k
SK	61.5k (9.79%)	2.2k	2.4k
TR	182k (29.07%)	4.0k	5.0k
RU	208k (33.21%)	4.8k	5.5k
PT	51.8k (8.25%)	1.2k	1.8k
CS	103k (16.42%)	3.5k	3.8k

Table 1: Numbers of sentences for LRLs and HRLs.

NMT Models. We experiment with Transformer models from the OpenNMT-py library version 3.4.3 (Klein et al., 2017).[6] All models are trained for 150k steps. The hyperparameter values are the default ones from the Transformer-Base model (Vaswani et al., 2017): 6 layers for the encoder and 6 for the decoder, 8 attention heads, label smoothing of 0.1, hidden layers with 512 units, and feed-forward networks with 2,048 units. The Adam optimizer (Kingma and Ba, 2014) is used. Following Atrio and Popescu-Belis (2022), we use a batch size of 8k tokens and the regularization parameters are: dropout rate of 0.3, scaling factor of 10, and gradients are re-normalized if their norm exceeds 5. In experiments with warm-up, there are 16k steps during which the learning rate increases from 0 to its maximum.

RL Agents. Several hyperparameters must be set for RL Agents. Their default values are given here, while the behavior of the systems when these values are modified are studied in Section 5.4.

[4] github.com/neulab/word-embeddings-for-nmt
[5] github.com/rsennrich/subword-nmt
[6] github.com/OpenNMT/OpenNMT-py

The TSCL algorithm is run with a smoothing coefficient $\alpha = 0.1$. The warm-up period is 16k steps, during which batches from HRLs are presented in a random order. For the ϵ-greedy policy, $\epsilon = 0.1$. These values correspond to those used by Matiisen et al. (2017).

The DQN algorithm is also run with a warm-up period of 16k steps on HRLs only. Unlike TSCL, a new action is selected every 10 steps, and not at every step, to reduce computing time, with no significant differences in observed results. The Q network underlying the RL agent has an input layer with 200 units, two fully connected subsequent layers with 512 units each, and an output layer with 8 units. As explained in Section 3.2, each value in the input layer corresponds to the cross-entropy loss of the NMT model over a batch of 10 sentences from a specific language.

The training of the Q network has a learning rate $lr = 2.5\text{e} - 4$ and a soft update smoothing coefficient $\tau = 0.005$. The discount factor, which influences the importance of future rewards in the agent's decision-making process, is $\gamma = 0.99$.[7] The experience replay buffer has minimal/maximal sizes of 1k/10k. These values are those used by Kumar et al. (2019).

Evaluation Metrics. Translation quality is measured using the BLEU and COMET metrics. BLEU scores are computed with the SacreBLEU library (Post, 2018).[8] COMET scores are computed using the `wmt22-comet-da` model (Rei et al., 2022).[9] Scores are computed using a rolling ensemble of four checkpoints. The best ensemble in terms of average BLEU score on the LRLs development sets is used to translate the test set.

5 Results and Analysis

5.1 Baselines

We compare TSCL and DQN to baseline training schedules in which source languages are selected randomly at each step, either with a uniform distribution ($P = 1/8$ for each language) or with a distribution that is proportional to the number of sentences of the respective language in the training data – hence between 0.95% for AZ and 33.21% for RU, as shown in Table 1. Moreover, a warm-up period of 16k on HRLs can be used or not. This results in four baseline schedules, shown in the first four lines of Tables 2, 3 and 4. While these baselines and the TSCL and DQN algorithms use monolingual batches, a fifth baseline uses multilingual shuffled ones, with sentences drawn randomly from the source languages in proportion to their frequency, and a warm-up period of 16k on HRLs. Shuffled batches were found to perform particularly well on this dataset (Neubig and Hu, 2018; Atrio et al., 2024).

5.2 Translation Performance

The BLEU and COMET scores of the TSCL and DQN algorithms, in comparison to the baselines, are presented in Table 2 for the LRLs and in Table 3 for the HRLs. The averages of BLEU and of COMET scores over the 8 languages are presented in Table 4, giving the same importance to each language, regardless of its frequency in the training data (macro-average).

The DQN algorithm outperforms on average all baselines, as well as the simpler TSCL algorithm, both in terms of BLEU and of COMET (Table 4). The TSCL algorithm is second for BLEU, but third for COMET, slightly behind the uniform training schedule with warm-up. Considering Table 2 with LRLs, we see that DQN often outperforms the other methods: it ranks first on COMET for AZ and BE and second for GL (but first on BLEU). Moreover, DQN ranks first on COMET for PT, as seen in Table 3. Therefore, DQN ranks first on three of the four least represented languages in the dataset.[10] This shows that DQN improves learning of the LRLs at the price of a small degradation in HRLs, though still improving their macro-average. As for TSCL, although it is competitive on average with the baselines, it lags behind the best ones when it comes to individual languages.

The baseline that is most often ranked first is the one that selects batches in proportion to the frequency of the language in the training data, with no warm-up. This has best BLEU and COMET scores on three HRLs (TR, RU, CS) and one LRL[10] (SK), likely because each of these languages constitutes more than 10% of the training data. However, in this case, the NMT model struggles to learn LRLs because it does not see enough data from them. As a result, when considering the macro-average, this

[7] This parameter is defined in Section 2 of Mnih et al. (2013) and is implicit in line 19 of our Algorithm 2.

[8] github.com/mjpost/sacrebleu, signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`.

[9] huggingface.co/Unbabel/wmt22-comet-da

[10] As noted in Section 4, the contrast between LRLs and HRLs made in previous work applies only within each pair of related languages.

Training schedule	Warm up	AZ → EN BLEU	AZ → EN COMET	BE → EN BLEU	BE → EN COMET	SK → EN BLEU	SK → EN COMET	GL → EN BLEU	GL → EN COMET
Uniform	no	13.86	62.99	19.75	60.80	32.85	75.09	31.14	72.08
Proportional	no	**15.82**	65.66	19.81	61.85	**35.11**	**76.51**	31.07	72.65
Uniform	16k	15.42	65.19	20.29	62.13	34.11	76.45	32.74	**73.84**
Proportional	16k	15.14	65.70	19.22	61.48	34.97	76.26	31.79	72.70
Shuffled batch	16k	14.37	64.37	20.08	62.15	33.92	76.28	32.15	72.99
TSCL	16k	14.89	65.04	20.10	61.96	34.35	76.23	32.64	73.59
DQN	16k	15.62	**65.86**	**21.11**	**62.82**	34.54	76.15	**33.02**	73.73

Table 2: Results of TSCL and DQN compared to baselines on LRLs.

Training schedule	Warm up	TR → EN BLEU	TR → EN COMET	RU → EN BLEU	RU → EN COMET	CS → EN BLEU	CS → EN COMET	PT → EN BLEU	PT → EN COMET
Uniform	no	27.26	75.41	26.95	72.18	30.87	74.30	39.76	78.41
Proportional	no	**29.40**	**77.48**	**28.14**	**74.04**	**32.47**	**75.79**	37.98	76.56
Uniform	16k	28.29	76.72	27.53	73.32	31.76	75.60	41.32	79.57
Proportional	16k	28.97	77.25	27.96	73.53	32.16	75.38	38.64	76.47
Shuffled batch	16k	28.25	76.76	27.31	73.20	31.30	75.37	40.58	79.25
TSCL	16k	28.50	76.64	27.56	72.99	31.76	75.15	**42.38**	79.52
DQN	16k	28.11	76.45	27.66	73.28	31.89	75.31	42.09	**79.73**

Table 3: Results of TSCL and DQN compared to baselines on HRLs.

Training schedule	Warm up	Average BLEU	Average COMET
Uniform	no	27.81	71.40
Proportional	no	28.73	72.57
Uniform	16k	28.93	72.85
Proportional	16k	28.61	72.35
Shuffled batch	16k	28.49	72.55
TSCL	16k	29.02	72.64
DQN	16k	**29.30**	**72.92**

Table 4: Macro-averages over all languages of the scores of TSCL and DQN compared to baselines.

baseline is slightly behind the one using warm-up on HRLs followed by selection of actions with uniform probability, which also has better COMET scores for BE, GL and PT.

Moreover, the DQN and TSCL algorithms are efficient in terms of convergence speed, defined as the number of steps needed to reach their best scores (the macro-average of BLEU on the LRLs of the development set). As shown more fully in Table 7 of Appendix A.2, DQN reaches best performance after 52k steps, followed closely by TSCL and by the baseline with proportional batches (both at 60k). The baseline with uniformly-drawn batches needs twice more steps to converge.

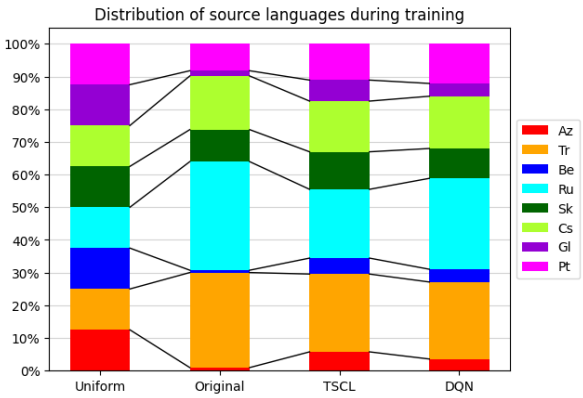


Figure 3: Proportions of data seen during training, as optimized by the TSCL and DQN algorithms, in comparison to uniform (1/8) or original proportions.

5.3 Optimized Training Schedule

We claim that the improved average scores with respect to the baselines are due to an optimized training schedule, which can be observed by considering the total amount of data from each lan-

guage seen during training, shown in Figure 3. The 'uniform' and 'proportional' baselines are shown in the first two columns. In the first case, the NMT model likely overfits to the LRL data, which is seen too often (12.5% of the times per language) with respect to its diversity (ca. 1% for three LRLs). In the second case, the number of times each LRL batch is seen during training is insufficient.

Our two algorithms strike a balance between these two extremes, as they are able to automatically determine more suitable proportions of batches of LRLs vs. HRLs for training. We see in Figure 3, third column, how the proportions of three LRLs are increased by TSCL (AZ in dark orange, BE in dark blue, and GL in dark purple). Two other languages with similar original proportions (PT in light purple and SK in dark green) see their proportions increased too, though less than the previous ones. Conversely, the proportions of HRLs decrease, especially for RU and TR.

In comparison to TSCL, the DQN algorithm appears to reach a slightly smaller proportion of LRLs, as seen in the fourth column of Figure 3, where proportions of the darker colors are shrunk with respect to the third column. These proportions are found quite quickly during training, as can be seen from Figure 6 in Appendix A.3, where we aggregate the proportions of actions every 1000 steps.

5.4 Role of Hyperparameters

In this section, we study the influence of hyperparameters on the scores of NMT systems trained with the TSCL and DQN algorithms. We present the scores obtained with significant variations of one parameter at a time in Table 5 for TSCL and in Table 6 for DQN. Globally, the scores of the algorithms do not vary much, which shows that they are robust with respect to the variations of the hyperparameters, but also confirms that the algorithms behave consistently from run to run. For both algorithms, BLEU and COMET scores lead to similar rankings.

For TSCL, we observe first that adding LRLs during warm-up (with uniform frequencies), or skipping warm-up entirely (thus starting with the highest learning rate), are not good options (second and third lines of Table 5). Instead, cross-lingual transfer from HRLs to LRLs becomes fully beneficial only with a 16k step warm-up on HRLs. Moreover, convergence is twice slower without warm-up. The smoothing coefficient α can vary around the default value of 0.3 with a small decrease in performance ($\alpha = 0.1$ is shown in the 4th line) and so can ϵ for the ϵ-greedy policy ($\epsilon = 0.3$ instead of 0.1 is shown in the 5th line). Finally, whether an action is selected every 10 steps or at every step results in comparable scores.

For DQN, we examine first if the Q network is over-parameterized, by reducing the size of the two hidden layers from 512 to 128 (2nd line of Table 6). This brings only a moderate decrease in average scores, but slightly better COMET scores for AZ, SK and GL.

If we vary τ, the smoothing rate of the updates of the Q network (see line 20 of Algorithm 2) within a large range between 0 and 1, the scores remain stable or even increase for some LRLs (3rd and 4th lines of Table 6, values of 0.5 and 0.995 with respect to default of 0.005).

Similarly, if we vary γ, the discount factor for the importance of future rewards, within a large interval between 0 and 1, the scores also remain stable (5th and 6th lines of Table 6, values of 0.5 and 0.01 with respect to default of 0.99). In fact, the value with the lowest scores is $\gamma = 0.01$, i.e. a system that gives only a marginal importance to long-term rewards. Conversely, this is also the system with the fastest convergence, although no particular variant seems to be particularly slow to converge (see Table 8 in Appendix A.2), and none achieves highest scores on all LRLs.

We can thus conclude that the default values of hyperparameters of TSCL and DQN used in Section 5.2 above, inspired respectively by Matiisen et al. (2017) and by Kumar et al. (2019), perform well and that both algorithms are stable when these hyperparameters vary.

5.5 Analysis of the Q Network

In this section, we propose a method to analyze the Q network of the DQN algorithm, which predicts on what language to train next, given a vector of 200 scores of an NMT model. Specifically, these scores are the cross-entropy loss values on 200 monolingual batches of 10 sentences each from the prototype set. At a given moment during training, the Q network can be probed with a specific vector as input, for instance a vector that represents a specific state of the NMT system. We propose to probe the Q network with a state in which one language is poorly learned. This is mimicked by assigning high loss values to the coefficients of the vector that represent scores on batches of this

Hyperparameter values	AZ → EN		BE → EN		SK → EN		GL → EN	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Default	**14.89**	**65.04**	**20.10**	**61.96**	**34.35**	**76.23**	**32.64**	**73.59**
Warm-up LRL+HRL	14.33	64.01	19.93	61.86	33.21	75.80	31.83	72.59
No warm-up	14.50	64.36	19.41	61.02	33.34	75.46	31.75	72.17
$\alpha = 0.3$	14.28	64.72	19.71	61.74	33.51	75.68	31.79	72.56
$\epsilon = 0.3$	14.10	64.37	19.95	61.80	33.36	75.75	31.16	72.38
$n = 10$	14.74	64.92	19.83	61.30	33.66	75.93	31.71	72.82

Table 5: MT performance using the TSCL algorithm when hyperparameters vary.

Hyperparameter values	AZ → EN		BE → EN		SK → EN		GL → EN	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Default	15.62	65.86	**21.11**	**62.82**	34.54	76.15	33.02	73.73
Hidden size = 128	15.86	**66.15**	20.38	62.64	34.40	76.21	32.59	73.79
$\tau = 0.5$	15.55	66.13	20.17	62.47	34.52	76.25	**33.13**	**73.81**
$\tau = 0.995$	**16.06**	66.06	20.19	62.18	**34.55**	**76.31**	32.45	73.46
$\gamma = 0.5$	15.78	65.85	20.51	62.64	34.49	76.16	32.94	73.63
$\gamma = 0.01$	15.34	65.38	20.59	62.36	33.87	75.70	32.43	73.39

Table 6: MT performance using the DQN algorithm when hyperparameters vary.

language. To avoid an entirely synthetic vector, we pick an actual vector occurring during training and multiply by 5 the loss values of all 25 batches from the targeted language.

We probe the Q network with each of the 8 source languages in turn, pretending that this language is not well learned and observing the action selected by the network, i.e. the language that it requires the NMT model to see next. Rather than observing the single selected language, we considered the softmaxed output activations for all 8 languages. The result is thus an 8-by-8 matrix, represented in Figure 4 at 28k and respectively 56k training steps. The X-axis represents the softmaxed output activations (predicted Q-values for each language), while the lines of the Y-axis correspond to each probed language (the one for which loss values were amplified).

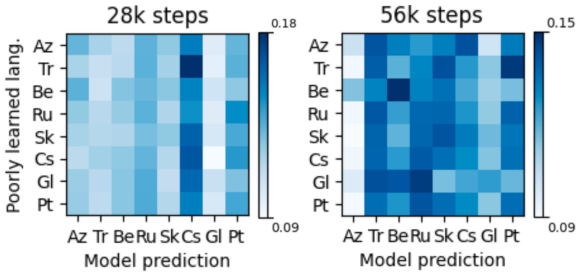


Figure 4: Behavior of the Q network at two stages during training.

The state at 28k steps is typical of incomplete training. The Q network appears to favor one HRL language (CZ) regardless of the language that is the least well learned according to its synthetic vector. The Q network selects one language for a large number of steps and gradually switches to another. The state at 56k steps (when the NMT trained with DQN reaches its best score) demonstrates a more balanced behavior: if one language is insufficiently learned, especially a LRL, then the network predicts that more training should be done on that language. Indeed, for several lines (though not all), the cell on the diagonal is one of the darkest of the line (e.g. for TR, BE, SK or PT). These observations suggest that the DQN model's decisions are complex and evolve over time, rather than always favoring the language that is currently the least well learned.

6 Conclusion and Perspectives

In this paper, we presented two algorithms for optimizing the training schedule of multilingual NMT models when a mixture of HRLs and LRLs must be learned on the source side. The TSCL algorithm models the expected return of each action by smoothing past observations, while DQN trains a neural network to perform this estimation and to select the optimal action.

Both algorithms strike a balance between a uniform distribution of training batches across lan-

guages and a distribution purely based on the respective frequencies of these languages in the actual data. The algorithms increase the proportions of LRLs and reduce those of HRLs, while still enabling cross-lingual transfer from HRLs to related LRLs. The better balance of HRLs and LRLs avoids too great a focus on the more abundant HRL data (which would sacrifice LRLs) or too great a focus on LRLs (which would lead to overfitting). Without such algorithms, it would be difficult to find extrinsic criteria to optimize the presentation frequencies of batches. Moreover, the optimized training schedules lead to improved macro-average BLEU and COMET scores.

We leave for future work the study of other ways to construct batches. One option is to use multilingual batches – though, as shown above, shuffled batches underperform with respect to an optimized balance of LRLs and HRLs. Another option is to define actions as specific batches or groups of batches, which would enable the model to prioritize certain batches over others, but would also increase the number of possible actions and hence the learning complexity of the RL agent.

The relevance of our proposal should be tested with additional datasets combining HRLs and related LRLs, and with other neural architectures for which cross-lingual transfer may be important to ensure acceptable performance on LRLs, particularly LLMs fine-tuned on translation tasks (Alves et al., 2024). In such cases, an optimized training schedule across available resources may also be beneficial.

Acknowledgments

We are grateful for its support to the Swiss National Science Foundation, through the DOMAT grant n. 175693, On-demand Knowledge for Document-level Machine Translation and the EXOMAT grant n. 228494, External Knowledge for Low-resource Machine Translation. We thank Giorgos Vernikos for feedback on earlier versions of the paper.

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.

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*.

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*.

Àlex R. Atrio, Alexis Allemann, Ljiljana Dolamic, and Andrei Popescu-Belis. 2024. Can the variation of model weights be used as a criterion for self-paced multilingual NMT? *arXiv 2410.04147*.

Àlex R. Atrio and Andrei Popescu-Belis. 2022. On the interaction of regularization factors in low-resource neural machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 111–120, Ghent, Belgium. European Association for Machine Translation.

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.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. 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 355–364, New Orleans, Louisiana. Association for Computational Linguistics.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model alignment as prospect theoretic optimization. *arXiv 2402.01306*.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. 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 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Peter J. Huber. 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101.

Sébastien Jean, Orhan Firat, and Melvin Johnson. 2019. Adaptive scheduling for multi-task learning. *arXiv 1909.06434*.

Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander Hauptmann. 2015. Self-paced curriculum learning. *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 1:2694–2700.

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.

Samuel Kiegeland and Julia Kreutzer. 2021. Revisiting the weaknesses of reinforcement learning for neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv 1412.6980*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Julia Kreutzer, David Vilar, and Artem Sokolov. 2021. Bandits don't follow rules: Balancing multi-facet machine translation with multi-armed bandits. *arXiv 2110.06997*.

Gaurav Kumar, George Foster, Colin Cherry, and Maxim Krikun. 2019. Reinforcement learning based curriculum optimization for 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 2054–2061, Minneapolis, Minnesota. Association for Computational Linguistics.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. *Advances in Neural Information Processing Systems (NeurIPS)*, 23.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436, Online. Association for Computational Linguistics.

Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. 2017. Teacher–student curriculum learning. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3732–3740, arXiv:1707.00183.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. 2013. Playing Atari with deep reinforcement learning. *arXiv 1312.5602*.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone. 2020. Curriculum learning for reinforcement learning domains: a framework and survey. *Journal of Machine Learning Research*, 21(1).

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

Long Ouyang et al. 2022. Training language models to follow instructions with human feedback. *arXiv 2203.02155*.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for 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 1162–1172, Minneapolis, Minnesota. 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.

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. *arXiv 2305.18290*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. *arXiv 1511.06732*.

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.

Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv 1609.04747*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv 1707.06347*.

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.

Yingli Shen and Xiaobing Zhao. 2024. Reinforcement learning in natural language processing: A survey. In *Proceedings of the 2023 6th International Conference on Machine Learning and Natural Language Processing*, MLNLP '23, page 84–90, New York, NY, USA. Association for Computing Machinery.

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.

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*, volume 30. Curran Associates, Inc.

Yu Wan, Baosong Yang, Derek F. Wong, Yikai Zhou, Lidia S. Chao, Haibo Zhang, and Boxing Chen. 2020. Self-paced learning for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1074–1080, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Chenglong Wang, Hang Zhou, Yimin Hu, Yifu Huo, Bei Li, Tongran Liu, Tong Xiao, and Jingbo Zhu. 2024. ESRL: Efficient sampling-based reinforcement learning for sequence generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19107–19115.

Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh. 2020a. Learning a Multi-Domain Curriculum for Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7711–7723, Online. Association for Computational Linguistics.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising Neural Machine Translation Training with Trusted Data and Online Data Selection. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 133–143, Brussels, Belgium. Association for Computational Linguistics.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576.

Xinyi Wang, Hieu Pham, Philip Arthur, and Graham Neubig. 2019. Multilingual neural machine translation with soft decoupled encoding. *arXiv 1902.03499*.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020b. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8526–8537, Online. Association for Computational Linguistics.

Jiawei Wu, Lei Li, and William Yang Wang. 2018a. Reinforced co-training. 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 1252–1262, New Orleans, Louisiana. Association for Computational Linguistics.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2018b. A study of reinforcement learning for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621, Brussels, Belgium. Association for Computational Linguistics.

Minghao Wu, Yitong Li, Meng Zhang, Liangyou Li, Gholamreza Haffari, and Qun Liu. 2021. Uncertainty-aware balancing for multilingual and multi-domain neural machine translation training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7291–7305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chen Xu, Bojie Hu, Yufan Jiang, Kai Feng, Zeyang Wang, Shen Huang, Qi Ju, Tong Xiao, and Jingbo Zhu. 2020. Dynamic curriculum learning for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3977–3989, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Asaf Yehudai, Leshem Choshen, Lior Fox, and Omri Abend. 2022. Reinforcement Learning with Large Action Spaces for Neural Machine Translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4544–4556, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Poorya Zaremoodi and Gholamreza Haffari. 2019. Adaptively scheduled multitask learning: The case of low-resource neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 177–186, Hong Kong. Association for Computational Linguistics.

Mingliang Zhang, Fandong Meng, Yunhai Tong, and Jie Zhou. 2021. Competence-based curriculum learning for multilingual machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2481–2493, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. An empirical exploration of curriculum learning for neural machine translation. *arXiv 1811.00739*.

Mingjun Zhao, Haijiang Wu, Di Niu, and Xiaoli Wang. 2020. Reinforced curriculum learning on pre-trained neural machine translation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34(05), pages 9652–9659.

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 Appendix

A.1 Epsilon Scheduling for DQN

The RL agent in the DQN algorithm follows, as explained in Section 3.2, an ϵ-greedy policy: with a probability of $1-\epsilon$, actions (i.e. the source language of a batch) are selected using the Q network, but with a probability of ϵ, a random action is selected. The exact schedule of ϵ is shown in Figure 5.

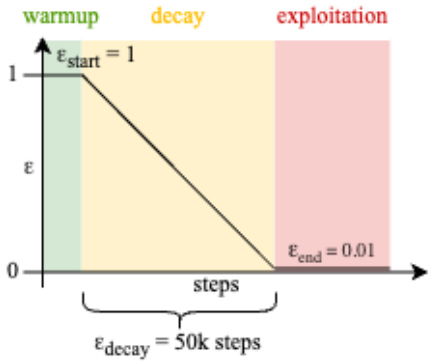


Figure 5: Evolution of ϵ during training with DQN.

During the warm-up period of 16k a value of 1 means that the Q network is not used and source HRLs (in this case) are drawn randomly. Then, during a decay of 50k steps, the importance of the Q network in deciding the actions grows progressively, while random choices decrease to a minimal probability of 0.01 after 66k steps. This approach, inspired by Kumar et al. (2019), achieves a balance between exploiting the Q network and exploring new actions.

A.2 Convergence Speed

In the experiments presented above, the scores were computed using a rolling ensemble of 4 checkpoints, and the best score was selected as the highest macro-average of BLEU achieved on the development data of the LRLs. We mentioned at the end of Section 5.2 that the DQN was the method that reached optimal scores after the smallest number of steps, followed by TSCL and then by the 'proportional' scheduling. The exact numbers of steps are given in Table 7, showing that DQN accelerates convergence with respect to the other schedules. Moreover, this behavior is stable when varying some of the hyperparameters of the algorithm, as shown in Table 8.

A.3 Learned Policies

We presented in Section 5.3 the total numbers of actions of each type (i.e. source language of the batch) selected during training for the TSCL and DQN algorithms, in comparison to the 'uniform' and 'proportional' training schedules. Here, we show in Figure 6 the evolution of the proportion of actions during training with the DQN algorithm, aggregated every 1000 steps.

In this representation, we first observe that the initial 16k steps are performed only on the HRLs, as configured. When the DQN algorithm starts playing a role, a random selection of languages is observed. As the algorithm learns, the proportions of LRLs decrease, while the proportions of HRLs increase, and tend to stabilize towards steady-state values. The proportions averaged over the entire training period are provided in the legend of the chart. These are the proportions compared between the systems in Figure 3.

Training	Warm up	Best checkpoint
Uniform	no	136k
Proportional	no	60k
Uniform	16k	124k
Proportional	16k	60k
Shuffle batch	16k	128k
TSCL	16k	56k
DQN	16k	52k

Table 7: Comparison of the number of steps required by the NMT model to achieve the best scores on the LRLs.

Parameter values	Best checkpoint
Default	52k
$\tau = 0.5$	60k
$\tau = 0.995$	76k
$\gamma = 0.5$	48k
$\gamma = 0.01$	36k
Hidden layer: 128	76k

Table 8: Comparison of the number of steps required for the NMT model using the DQN algorithm to achieve the best scores on the LRLs. The parameter values are the default ones, except the changes shown for each line.

A.4 The TSCL Algorithm

The full specification of the TSCL algorithm in pseudo-code is provided hereafter as Algorithm 1.

A.5 The DQN Algorithm

The full specification of the DQN algorithm in pseudo-code is provided hereafter as Algorithm 2.

Figure 6: Evolution of the proportions of actions (i.e. language of batches) chosen during the training of the DQN system. Aggregation windows of 1000 steps are computed. The first 16k steps are the warmup on the HRLs only.

Algorithm 1: TSCL algorithm for NMT.

Require: actions $\mathcal{A} \leftarrow \{A_1, .., A_k\}$, number of training steps ts, number of consecutive actions n, number of warm-up steps w, ϵ-greedy policy exploration parameter ϵ, smoothing coefficient α

Initialize NMT model
Initialize action index $i \leftarrow 1$
Initialize unvisited actions indexes $U \leftarrow \{2, .., k\}$
Initialize estimated return $Q(A_k) \leftarrow 0$ for all k actions
Initialize rewards history $H(A_k) \leftarrow 0$ for all k actions
for $t \leftarrow 1, \ldots, ts$ **do**
    Sample mini-batch B_t from action A_i
    Train NMT model using mini-batch B_t
    if $t \bmod n = 0$ **then**
        Observe reward $R_t \leftarrow X_t - H(A_i)$
        Update reward history $H(A_i) \leftarrow X_t$
        Exponentially smooth estimated return $Q(A_i) \leftarrow \alpha R_t + (1 - \alpha) Q(A_i)$
        if $|U| \neq 0$ **then**
            Choose action index $i \leftarrow U[0]$
            Update $U \leftarrow U - \{i\}$
        end
        else
            Choose random number r between 0 and 1
            if $t < w$ ***or*** $r < \epsilon$ **then**
                Choose action index i randomly between 1 and k
            end
            else
                Set i as the index of the max arg. of absolute values in Q
            end
        end
    end
end

Algorithm 2: DQN Algorithm for NMT

Require: actions $\mathcal{A} \leftarrow \{A_1, .., A_k\}$, number of training steps ts, number of consecutive actions n, number of warm-up steps w, ϵ-greedy policy exploration parameter ϵ, soft update coefficient τ, replay memory capacity c, minimum replay memory capacity c_{min}

Initialize NMT model learning algorithm
Initialize RL agent's online model
Initialize replay memory deque $\mathcal{D}$ with capacity c
Initialize RL agent's target network with same weights as RL agent's online model
Initialize action index $i \leftarrow 1$
for $t \leftarrow 1, \ldots, T$ **do**
    Sample mini-batch B_t from action A_i
    Train NMT model using mini-batch B_t
    if $t \bmod n = 0$ **then**
        if $t < w$ **then**
            Choose action index i randomly between 1 and k
        end
        else
            Observe current state S_t
            Observe reward $R_t \leftarrow X_t - X_{t-1}$
            Store transition (S_{t-1}, i, R_t, S_t) in replay memory $\mathcal{D}$
            if $|D| >= c_{min}$ **then**
                Sample mini-batch of transitions T from replay memory $\mathcal{D}$
                Train RL agent's online model using mini-batch T
                Soft update RL agent's target model weights with RL agent's online model weights $\theta^- \leftarrow \tau\theta + (1-\tau)\theta^-$
            end
            Choose random number r between 0 and 1
            if $r < \epsilon$ **then**
                Choose action index i randomly between 1 and k
            end
            else
                Predict Q values at state S_t with RL agent target network
                Set i as the index of the arg. max in Q
            end
            Decrease ϵ according to decay schedule
        end
    end
end

Languages Transferred Within the Encoder: On Representation Transfer in Zero-Shot Multilingual Translation

Zhi Qu[*†] **Chenchen Ding**[†‡] **Taro Watanabe**[†]

[†]Nara Institute of Science and Technology, Japan
{qu.zhi.pv5, taro}@is.naist.jp

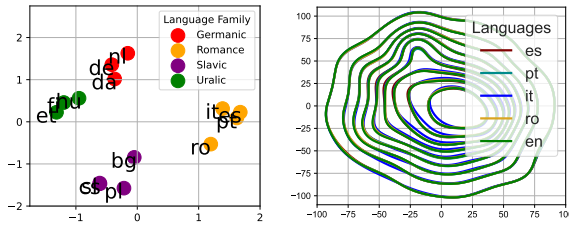
[‡]National Institute of Information and Communications Technology, Japan
chenchen.ding@nict.go.jp

Abstract

Understanding representation transfer in multilingual neural machine translation (MNMT) can reveal the reason for the zero-shot translation deficiency. In this work, we systematically analyze the representational issue of MNMT models. We first introduce the identity pair, translating a sentence to itself, to address the lack of the base measure in multilingual investigations, as the identity pair can reflect the representation of a language within the model. Then, we demonstrate that the encoder transfers the source language to the representational subspace of the target language instead of the language-agnostic state. Thus, the zero-shot translation deficiency arises because the representation of a translation is entangled with other languages and not transferred to the target language effectively. Based on our findings, we propose two methods: 1) low-rank language-specific embedding at the encoder, and 2) language-specific contrastive learning of the representation at the decoder. The experimental results on Europarl-15, TED-19, and OPUS-100 datasets show that our methods substantially enhance the performance of zero-shot translations without sacrifices in supervised directions by improving language transfer capacity, thereby providing practical evidence to support our conclusions. Codes are available at https://github.com/zhiqu22/ZeroTrans.

1 Introduction

State-of-the-art neural machine translation systems are adaptable to multilingualism, resulting in a single encoder-decoder model that executes arbitrary translations by adding a tag specified to the target language at the beginning of source sentence (Firat et al., 2016; Johnson et al., 2017; Wu et al., 2021). Multilingual neural machine translation (MNMT) is theoretically attractive because zero-shot translations, i.e., translations unseen in training, allow the training of a multilingual model with minimal cost. Unfortunately, the performance of zero-shot translations always lags behind (Aharoni et al., 2019; Arivazhagan et al., 2019a; Gu et al., 2019; Yang et al., 2021; Pan et al., 2021; Chen et al., 2023a).



(a) Cluster by target languages (b) Semantic alignments

Figure 1: Different analytical methods lead to different conclusions. 1a means the target language family clusters the representations of translations from English (en) to other languages through the encoder. 1b indicates the encoder semantically aligns different source languages. Language codes in this work follow ISO 639-1, and Appendix D provides details of those figures.

Representational analysis in MNMT models can guide the improvement of zero-shot translation. However, two contrary opinions are demonstrated by prior works: (1) the encoder clusters translation representations based on the target language (Kudugunta et al., 2019; Liu et al., 2021; Tan and Monz, 2023; Stap et al., 2023; Sun et al., 2024), as illustrated in Figure 1a; (2) an ideal encoder is expected to learn language-agnostic representations, capturing general cross-lingual features that are transferable across languages (Pan et al., 2021; Gu and Feng, 2022; Gao et al., 2023; Bu et al., 2024), as shown in Figure 1b. In this work, we aim to analyze and reconcile this discrepancy. We first introduce the identity pair, a pseudo pair translating a sentence to itself. Specifically, the analyses con-

[*]This work was done during the first author's internship at Advanced Speech Translation Research and Development Promotion Center, National Institute of Information and Communications Technology, Kyoto, Japan.

ducted by those prior works rely on real translation pairs, leading to inaccurate results, as a translation pair cannot serve as a base measure for another pair. The identity pair, however, addresses this issue by serving as a proxy for the optimal representation of a language instead of a translation pair. Then, multiple analytical methods are employed to show the representation transfer within MNMT models. Our findings offer a unified perspective on these two opinions: the encoder transfers translation representations into the target language subspace, where different source languages are semantically aligned. Thus, the zero-shot translation deficiency stems from the failure to transfer the translation representation to the target language, as it becomes entangled with representations of other languages in the encoder.

Guided by our findings, we propose two methods for the encoder and decoder, respectively, to improve multilingual representations: **Lo**w-Rank **L**anguage-specific **E**mbedding (LoLE) is applied to bias the representations in the subspaces of target languages at the encoder; and **L**anguage-specific **C**ontrastive **L**earning of **R**epresentations (LCLR) is applied at the decoder to isolate representational space across languages. We evaluated the proposed methods on three benchmarks, Europarl-15 (Koehn et al., 2005), TED-19 (Ye et al., 2018), and OPUS-100 (Zhang et al., 2020a; Yang et al., 2021), for two automatic metrics, SacreBLEU (Post, 2018) and BERTScore (Zhang et al., 2020b). The experimental results show that our methods outperform strong baselines in training from scratch because of improved representational transferability. Our methods also perform effectively in fine-tuning, even though pre-trained models are trained by different strategies of language tags, which proves that target language information on the encoder side consistently benefits MNMT.

2 Background

2.1 Multilingual Neural Machine Translation

Johnson et al. (2017); Wu et al. (2021) demonstrated that the training strategy of adding a language tag at the beginning of the input sentence on the encoder side boosts the zero-shot translation capacity of the MNMT model. Given a multilingual corpus $\mathbb{C}$ that covers a set of t languages, a set of their corresponding language tags exists: $\mathbb{L} = \{l_1, l_2, \ldots, l_t\}$. For a source-target sentence pair $(\boldsymbol{x}, \boldsymbol{y})$, i.e., $\boldsymbol{x} = x_1, x_2, \ldots, x_n$ and $\boldsymbol{y} = y_1, y_2, \ldots, y_m$, the training data consists of a pair in form of $(\boldsymbol{x}, l, \boldsymbol{y})$, where l is the language tag of $\boldsymbol{y}$ that instructs translation from $\boldsymbol{x}$ into language l. The model is trained over all pairs in $\mathbb{C}$ to optimize the following cross-entropy loss:

$$\mathcal{L}_{ce} = -\sum_{\boldsymbol{x},l,\boldsymbol{y}\in\mathbb{C}} \log p(\boldsymbol{y}|l, \boldsymbol{x}; \theta), \qquad (1)$$

where $p(\boldsymbol{y}|l, \boldsymbol{x}; \theta)$ is the probability distribution of $\boldsymbol{y}$ and θ represents the model parameters.

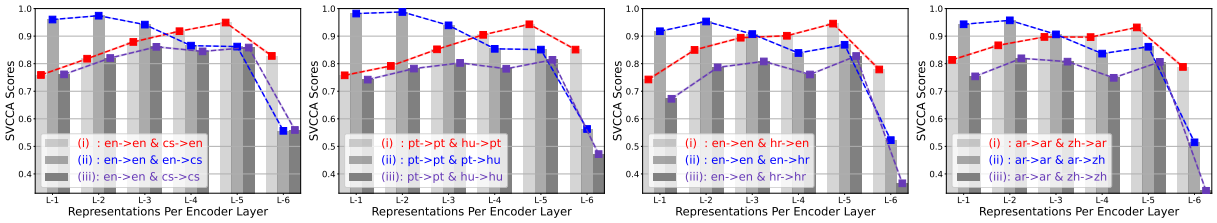
2.2 The Discrepancy in Prior Works

Pan et al. (2021); Gao et al. (2023); Bu et al. (2024) state that, for an encoder-decoder MNMT model, an ideal encoder is regarded as transferring the source sentence into a language-agnostic state, preserving only semantic information.[1] As evidence, the t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008), which can convert similarities between vectors into joint probabilities, has been used to show that representations of sentences from different languages are aligned at the output of the encoder when sharing the same semantics. However, this result contrasts with the findings of Kudugunta et al. (2019); Stap et al. (2023); Tan and Monz (2023). Specifically, using the singular value canonical correlation analysis (SVCCA) (Raghu et al., 2017) to compare the similarity between two vectors, i.e., the sentence representations of two translations, reveals that the encoder tends to transfer the representation into a state with target language features.

We argue that this discrepancy stems from the lack of a base measure. Namely, those works always compare the representations of real translation pairs in which different analysis methods lead to different results. For instance, the translation from English to German, denoted by en→de, cannot be accurately measured by comparing it with another translation from a different language x→de, because en→de is expected to be measured by the language representation of either de or en. Thus, proposing a base measure is necessary to draw the same conclusion from different analysis methods, e.g., t-SNE or SVCCA[2].

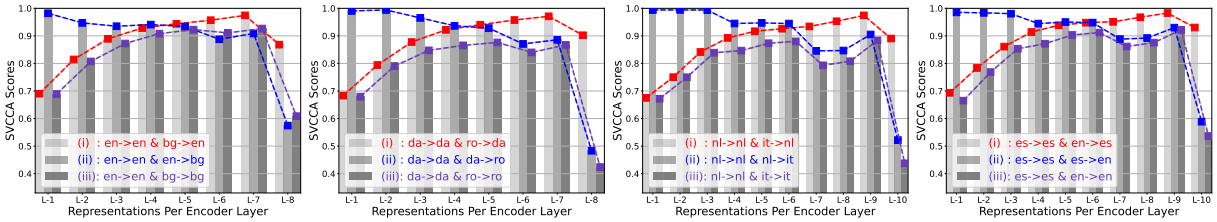
[1] Although Pan et al. (2021) proposed that the ideal output of the encoder is language-agnostic by adding a source language tag at the beginning of the encoder, the follow-up works (Gao et al., 2023; Bu et al., 2024) practiced this concept with adding a target language tag, which is the main strategy investigated in this work.

[2] We follow Liu et al. (2021) to measure SVCCA scores at the sentence level, which is introduced in Appendix A.



(a) (en, cs) in Europarl. (b) (pt, hu) in Europarl. (c) (en, hr) in TED. (d) (ar, zh) in TED.

Figure 2: Visualizations of layer-wise SVCCA scores for the encoder. (①, ②) indicate the source language and target language, respectively. The analyzed models have 6 encoder layers, and the analysis based on models with 8 and 10 encoder layers is shown in Figure 3.



(a) (en, bg) with 8 layers. (b) (da, ro) with 8 layers. (c) (nl, it) with 10 layers. (d) (es, en) with 10 layers.

Figure 3: Visualizations of layer-wise SVCCA scores for the encoders with 8 and 10 layers on diverse languages in Europarl-15, as a comparison of Figure 2 to prove the generalization.

3 Investigating Representation Transfer in MNMT

We conduct preliminary experiments to investigate representations by introducing identity pairs as base measures using two different datasets, Europarl-15 (Koehn et al., 2005) and TED-19 (Ye et al., 2018), which are introduced in Appendix B in detail. Then, following Kudugunta et al. (2019), our investigation is based on Transformer models with 6 encoder and decoder layers. We also investigate scenarios with 8 and 10 encoder layers. Appendix C introduces the detailed model settings.

3.1 Identity Pairs

An identity pair refers to a sentence pair translating from one sentence into itself to represent the optimal state of processing language features, i.e., the semantics and syntax of the source sentence by the model. Notably, our models are only trained by translating from one language to another. In this setup, the identity pair is a zero-shot translation, which does not simply copy the input to the output.[3] On the encoder side, we derive the representation from a language translating to itself, i.e., $(\boldsymbol{x}, l', \boldsymbol{x})$, where l' is the language tag of $\boldsymbol{x}$, with the aim of recovering the source sentence from the hidden representations without inference on the decoder side. We also derive the representation in the decoder from the gold translation of $(\boldsymbol{x}, l', \boldsymbol{x})$.

We use SacreBLEU (Papineni et al., 2002; Post, 2018) to evaluate the translation quality of 6 identity pairs, which are generated by inference. The scores of en→en, de→de, and pt→pt in Europarl-15 are 73.49, 61.04, and 71.97, which significantly outperform 44.04 of de→en, 36.63 of en→de, and 46.24 of en→pt, respectively. Similarly, en→en, tr→tr, and vi→vi in TED-19 obtain scores of 72.52, 36.58 and 59.26, which are higher than 34.92 of de→en, 14.81 of en→tr, and 29.78 of en→vi, respectively. Such high scores in the identity pair are caused by that short sentences are recovered from hidden representations perfectly, and long sentences only have a few changes in word selection. Such evidence suggests that identity pairs can serve as base measures for comparing representations because the identity pair is a proxy for the optimal representation of a language, specifically, x→en are expected to be close to en→en in the representational space.

3.2 Language Transfer Within the Encoder

Given two languages ① and ②, we follow Pan et al. (2021); Liu et al. (2021) to obtain sentence-level representations for $\boldsymbol{x}$ in ① and $\boldsymbol{y}$ in ② by applying mean pooling over token representations. We then

[3]This claim is supported by Qu and Watanabe (2022), which demonstrate another zero-shot scenario: removing the language tag during inference results in any source sentence being translated into English. Thus, the identity pair indeed presents a translation process by adding a language tag.

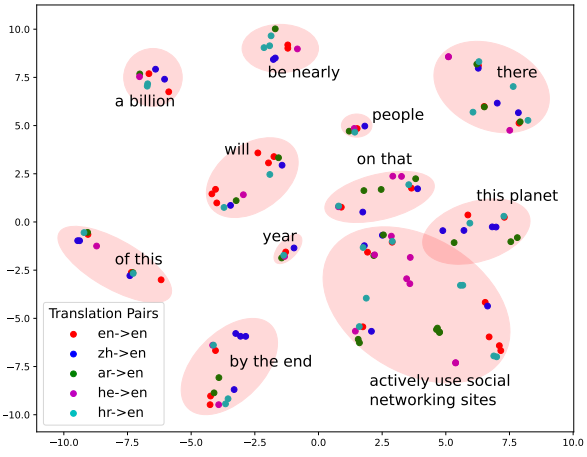
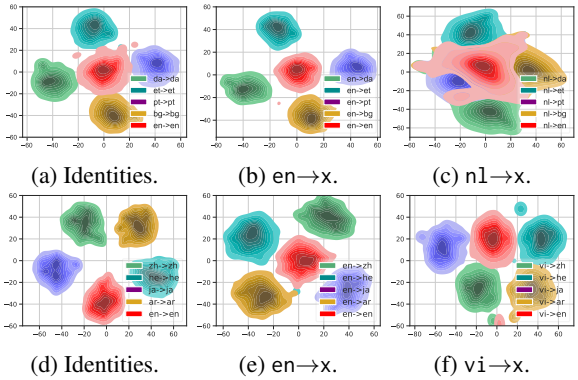


Figure 4: t-SNE plot of the token-level alignment between en→en and x→en in TED-19. Each point is a token's representation collected from the output of the encoder. Representations of different tokens are clustered by the semantics, which are denoted by English phrases, where the overall variance is 0.09. Appendix G shows the more details.



(a) Identities. (b) en→x. (c) nl→x.

(d) Identities. (e) en→x. (f) vi→x.

Figure 5: Visualizations for the encoder's output by t-SNE and BiKDE. 5a, 5b and 5c are measured in Europarl-15. 5d, 5e and 5f are measured in TED-19.

organize the comparisons into three cases to analyze the variation in encoder representations: (i) comparing $(\boldsymbol{x}, l^{①}, \boldsymbol{x})$ and $(\boldsymbol{y}, l^{①}, \boldsymbol{x})$ to show how target language features are encoded; (ii) comparing $(\boldsymbol{x}, l^{①}, \boldsymbol{x})$ and $(\boldsymbol{x}, l^{②}, \boldsymbol{y})$ to show how source language features are encoded; (iii) comparing two different identities, $(\boldsymbol{x}, l^{①}, \boldsymbol{x})$ and $(\boldsymbol{y}, l^{②}, \boldsymbol{y})$[4].

The two models trained by Europarl-15 and TED-19 show the same tendency in Figure 2, i.e., the language features for ① of (i) consistently increase in both cases involving the central language, i.e., English, in Figures 2a and 2c, and non-central languages in Figures 2b and 2d. The target language feature of ① emerges as the primary factor that affects representations at the fifth and sixth layers when the cases of (i), (ii), and (iii) are compared. Therefore, we can conclude that the language features of the representations are transferred to the target side within the encoder. Meanwhile, we observe that the scores of (iii) are close to or even exceed those of (ii) at some layers both in Figure 2. This proves that the feature of the source language is not the primary factor for encoding representations because representations are transferred to the subspaces of target languages. Thus, the comparison between (ii) and (iii) supports that language transfer is completed within the encoder.

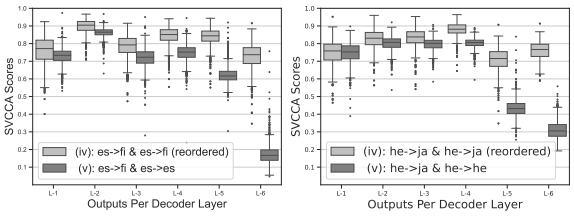
To validate the generality of this conclusion, we extend our analysis to models with 8 and 10 encoder layers (Figure 3). The same trends hold: (i) continues to show increasing similarity scores, with the final values even higher than in the 6-layer setting, suggesting stronger target language alignment. Again, the relationship between (ii) and (iii) remains consistent, further confirming that the source language is not the dominant factor in shaping encoder representations. These results demonstrate that language transfer within the encoder is robust across different architectural depths.

On the other hand, identity pairs also allow the measurement of the alignment of different languages in the target language space through t-SNE. Compared with the sentence-level measurement of Pan et al. (2021); Gao et al. (2023), we measure the alignment of representations at the token level. As shown in Figure 4, semantic similarity causes the representations to cluster together. Moreover, as shown in Appendix G, these representations are not clustered before being processed by the encoder, and the case with different target languages has a higher overall variance. Combined with the finding that the encoder transfers the representation of the source language to the target language, the evidence further suggests that there is no general and cross-lingual state for directly sharing semantic information within the encoder, and the alignment shown in Figure 1b occurs in the representational subspace of the target language.

3.3 Entanglements Hindering the Transfer

Although the investigation in Section 3.2 shows that the representations gradually transfer to the target language in a translation pair, the representation spaces of multiple languages may potentially entan-

[4] $(\boldsymbol{x}, l^{①}, \boldsymbol{x})$ indicates the identity of ①, i.e., ①→①, and $(\boldsymbol{y}, l^{①}, \boldsymbol{x})$ indicates a sentence of ② translating to the sentence of ① instructed by the language tag of ①, i.e., ②→①.



(a) (es, fi) in Europarl. (b) (he, ja) in TED.

Figure 6: Visualizations of layer-wise SVCCA scores for the decoder. (③, ④) shows the involved languages.

gle with each other, resulting in the failure of the zero-shot translation (Qu and Watanabe, 2022). To further illustrate the relationship between different languages, we use t-SNE and BiKDE to visualize the representations at the output of the encoder for the several identity pairs in Europarl-15 and TED-19. Figures 5a and 5d show that different identity pairs are uniformly distributed in the representational space. This distribution proves again that the encoder is language-specific because each language has an isolated representational subspace.

Compared with identity pairs that represent the ideal capability of the model in processing languages, the distributions plotted in Figures 5b and 5e reflect the actual capacity for the supervised translation of en→x. Figures 5b and 5e show the distribution of representations in the pairs translating from en, which are similar to that of identity pairs. The difference between identity pairs and supervised language pairs can be attributed to the influence of the source language information, which hinders the full use of the target language information learned by the encoder.

Moreover, the language-specific subspaces cannot be clearly separated for zero-shot translations, as shown in Figures 5c and 5f. Specifically, all representations are entangled around the supervised language pair of x→en, which hinders these representations from transferring into the ideal subspaces of the target language. This aligns with Qu and Watanabe (2022) and Stap et al. (2023) that multilingual representations are entangled, which explains the weakness of zero-shot translation compared with supervised translation, suggesting that improving the transferability of representations is attributed to the extent of language transfer within the encoder.

3.4 Language Features in the Decoder

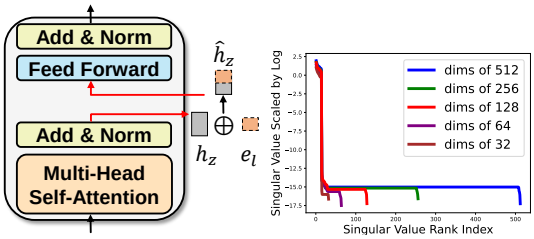
We further investigate the importance of target language features versus semantics in the decoder. Given two sentences $\boldsymbol{x}$ of language ③ and $\boldsymbol{y}$ of language ④, the decoder representation of $(\boldsymbol{x}, l^{④}, \boldsymbol{y})$ is considered as the base measure. We group two cases: (iv) For each sentence in a test set, we identify the pair $(\boldsymbol{x}', l^{④}, \boldsymbol{y}')$ with the lowest SVCCA score in the encoder representation to derive a $\boldsymbol{x}'$ that is distant from $\boldsymbol{x}$. Then, we compare it with the base measure to show the importance of target language features; (v) We compare the base measure and $(\boldsymbol{y}, l^{④}, \boldsymbol{y})$ to show the importance of semantics. The two scenarios shown in Figure 6 present the same trend, which is that (iv) maintains high scores despite their semantics being entirely different. At the top layers of the decoder, the gradually increasing difference between (iv) and (v) confirms that the decoder tends to learn the target language specificity (Sen et al., 2019). However, Figure 6 shows that, for (iv), a wider interquartile range exists at the bottom layers of the decoder, and its scores are close to those of (v), which implies the weakness in distinguishing languages for zero-shot translations.

4 Encouraging Representation Transfer

To validate the findings in Section 3, we propose two methods on the encoder and decoder sides, respectively, to improve transferability. Based on the findings in Sections 3.2 and 3.3, improving the extent of language transfer in the encoder can overcome the hindered representations of zero-shot language pairs. We introduce a learnable embedding referred to as Low-rank Language-specific Embedding (LoLE). It serves as biases to force representations to transfer into the target language with negligible cost. Based on the findings in Section 3.4, the capacity for multilingual features is insufficient at the lower layers of the decoder. We introduce Language-specific Contrastive Learning of Representations (LCLR) as an training extra task to regularize the representations to specify the representational boundary for each language.

4.1 Low-Rank Embedding for the Encoder

Let $\mathbb{E} = \{\boldsymbol{e}^1, \boldsymbol{e}^2, \ldots, \boldsymbol{e}^p\}$, $\boldsymbol{e}^j \in \mathbb{R}^d$, be a set of embeddings that correspond one-to-one with the languages in $\mathbb{L}$. For a translation $(\boldsymbol{x}, l, \boldsymbol{y})$, the embedding in $\mathbb{E}$ corresponding to l is denoted by $\boldsymbol{e}^l$. The hidden representation $\mathbf{H}^z = \{\boldsymbol{h}_1^z, \boldsymbol{h}_2^z, \ldots, \boldsymbol{h}_q^z\}$, where $\mathbf{H}^z \in \mathbb{R}^{q \times d}$, is extracted before the feed-forward network (FFN) (Khandelwal et al., 2021; Xu et al., 2023; Deguchi et al., 2023) at the z-th encoder layer. Then, we broadcast $\boldsymbol{e}^l$ to $\mathbf{E}^l$,



(a) Process (b) Dimension

Figure 7: Illustrations of LoLE. 7a shows the process in an encoder layer, where $\bigoplus$ represents the operation of Equation(2) at a low rank. 7b is the spectrum with different dimensions of $\mathbb{E}$, which illustrates the singular values of the covariance matrix of $\mathbb{E}$ in sorted order and logarithmic scale.

$\mathbf{E}^l \in \mathbb{R}^{q \times d}$, and we bias $\mathbf{H}^z$ to $\hat{\mathbf{H}}^z$:

$$\hat{\boldsymbol{h}}_i^z = \boldsymbol{h}_i^z + \boldsymbol{e}_i^l, \tag{2}$$

where $\hat{\mathbf{H}}^z$ is the input for the FFN of the z-th encoder layer (Figure 7a). We execute this biasing at the second-top encoder layer to ensure sufficient capacity for fusing representations and language information, while implicitly allowing lower layers to focus on surface-level information.

The simple language categorization by embedding may lead to a risk of dimensional collapse in the latent space (Jing et al., 2022). Thus, we reduce the dimension of $\mathbb{E}$ to d^e to allow biasing in a low rank, and add it to the head of $\boldsymbol{h}_i^z$ to simultaneously encourage language transfer and minimize the influence on representations (Hu et al., 2021). Figure 7b is a spectrum used to illustrate dimensional collapse using a comparison of different d^e in Europarl-15. The spectrum shows that larger dimensions are primarily composed of noise, whereas a dimension that is too small adversely affects the learning of key features.

4.2 Contrastive Learning for the Decoder

Given a training batch, we extract hidden representations from the output of each decoder layer and apply averaged pooling to obtain a fixed-dimensional representation for each sentence. To avoid dimensional collapse (Tian, 2022; Jing et al., 2022), we also use the head of the representation for contrastive learning, i.e., the vectors in the batch $\mathbb{B} = \{\overline{\boldsymbol{h}}_1, \overline{\boldsymbol{h}}_2, \dots\}, \overline{\boldsymbol{h}}_i \in \mathbb{R}^{d^h}, d^h < d$.

To prevent a potential invalid training objective in sampling caused by the skewed distribution in a batch, we first define $\mathbb{B}' \subseteq \mathbb{B}$ by omitting instances that do not share their target language with any other instance in $\mathbb{B}$. For a given instance of $\boldsymbol{h}^{\text{anc}} \in \mathbb{B}'$, which is the anchor in contrastive learning, we let $\mathbb{B}^+$ denote the subset of $\mathbb{B}'$, including instances with the same target language as $\boldsymbol{h}^{\text{anc}}$, where $|\mathbb{B}^+| > 1$. Likewise, we define a subset for negative instance $\mathbb{B}^- = \mathbb{B}' \setminus \mathbb{B}^+$. For contrastive learning, we randomly sample the positive instance $\boldsymbol{h}^{\text{pos}}$ from $\mathbb{B}^+$ and sample k negative instances $\boldsymbol{h}^{\text{neg}}$ from $\mathbb{B}^-$. Additionally, if $k > |\mathbb{B}^-|$, we dynamically clip k to $|\mathbb{B}^-|$. Formally, the objective of LCLR is formulated as

$$\begin{aligned} \mathcal{L}_{ctr} &= -\sum_{\boldsymbol{h}^{\text{anc}} \in \mathbb{B}'} \log \frac{e^{s^+}}{e^{s^+} + \sum_{i=1}^{k} e^{s_i^-}}, \\ s^+ &= \text{sim}(\boldsymbol{h}^{\text{anc}}, \boldsymbol{h}^{\text{pos}}), \boldsymbol{h}^{\text{pos}} \in \mathbb{B}^+, \\ s_i^- &= \text{sim}(\boldsymbol{h}^{\text{anc}}, \boldsymbol{h}_i^{\text{neg}}), \boldsymbol{h}_i^{\text{neg}} \in \mathbb{B}^-, \end{aligned} \tag{3}$$

where $\text{sim}(\cdot)$ calculates the similarity of representations using the cosine similarity. The final training objective is the sum of Equations 1 and 2, i.e.,

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{ctr}. \tag{4}$$

5 Experiments

5.1 Setup

Datasets Our experiments comprise three popular English-centric datasets, i.e., the training and validation sets only involving translation pairs translating to en or from en, including Europarl-15 (Koehn et al., 2005; Dabre and Kurohashi, 2019), TED-19 (Ye et al., 2018) and OPUS-100 (Zhang et al., 2020a; Yang et al., 2021). The details of those datasets can be found in Appendix B.

Evaluation We evaluate the performance of models on the test sets of those three datasets and set the beam size to 4 in inference. We employ SacreBLEU (Papineni et al., 2002; Post, 2018) to evaluate the quality of inferences at the word level and report BERTScore (Zhang et al., 2020b) of inferences at the representation level. We measure the off-target ratio on zero-shot translations as a supplement. We also conduct statistical significance testing (Koehn, 2004). We describe our motivation in selecting evaluation metrics, the evaluation details, and the implementation of statistical significance testing in Appendix H.

Models When training from scratch, we implement a Transformer model with 6 encoder and decoder layers. Given that those three datasets have different sizes, we set different hyper-parameters in

	Europarl-15						TED-19						OPUS-100			
	BLEU			B.S.			BLEU			B.S.			BLEU			B.S.
Method	en→	→en	zero.	en→	→en	zero.	en→	→en	zero.	en→	→en	zero.	en→	→en	zero.	zero.
VANILLA	37.49	43.39	24.65	88.50	95.71	84.27	**24.53**	29.67	11.98	**83.77**	93.54	77.74	23.37	28.30	5.04	69.98
DISPOS	37.15	43.37	25.89	88.39	**95.72**	84.69	24.08	29.43	12.80	83.62	93.49	78.36	22.72	28.24	5.58	70.74
TLP	37.41	43.28	24.96	88.47	95.71	84.40	24.44	29.62	12.74	83.73	93.53	78.24	**23.41**	28.30	4.60	69.40
SEMALI	37.27	43.06	25.25	88.42	95.69	84.43	23.55	28.67	**13.45**†	83.43	93.36	**78.91**†	22.35	28.29	6.42	72.00
LOLE	37.62	43.50	26.09†	**88.51**	**95.72**	84.81	24.39	29.72	13.20	83.74	93.54	78.65	23.15	28.28	7.92†	**73.32**†
LCLR	37.44	43.43	25.71	88.46	**95.72**	84.66	24.46	29.66	12.12	83.76	93.54	77.87	23.34	**28.37**	5.11	70.04
BOTH	**37.67**	**43.51**	**26.20**†	88.50	**95.72**	**84.85**†	24.49	**29.79**	13.31†	83.76	**93.56**	78.76†	23.40	28.27	**7.97**†	73.10†

Table 1: Averaged scores for experiments of training from scratch. BOTH means using LOLE and LCLR together; en→ and →en abbreviates en→x and x→en; zero. means zero-shot language pairs; and B.S. abbreviates BERTScore. We only report zero-shot language pairs of OPUS-100 because BERTScore does not support some pairs in supervised translations, but zero-shot translation pairs of OPUS-100 are involved only with 6 languages, which are supported. The bold number indicates the best result and the numbers with † are significantly better than VANILLA according to the significance test with $p < 0.05$. The off-target ratios are reported in Appendix I.

training. Then, three open-source models[5] are utilized in fine-tuning experiments, including M2M-418M, M2M-1.2B (Fan et al., 2020) and mBART50 (Tang et al., 2020). The hyper-parameter settings can be found in Appendix C. Additionally, hyper-parameters are selected based on the ablation studies conducted on the validation sets, which is reported in Appendix E.

Baselines Vanilla Transformer (Vaswani et al., 2017; Johnson et al., 2017) is one of the baselines, denoted by VANILLA in the experiments of training from scratch. Then, the baseline in fine-tuning experiments is the full-parameter fine-tuning, denoted by F.T.. Moreover, three representative methods are reproduced in our experiments of training from scratch, the standard of baseline selection is as follows:

- SEMALI: Pan et al. (2021) think the encoder output is language-agnostic, so they align the semantic information across different languages at the output of the encoder. However, our analysis shows that this viewpoint is inaccurate because the semantic information is aligned by the subspace of the target language instead of the real language-agnostic. When there are not any additional parameters introduced, SEMALI still is the de-facto SOTA based on regularizing representations in MNMT.

- DISPOS: Liu et al. (2021) have the same objective as LoLE, however, they suggest reducing the constraint on the encoder (Gu et al., 2019) by removing the residual connection, which is a different style that corresponds to the idea of biasing we used in LoLE.

- TLP: Yang et al. (2021) aim to add a loss to predict the language id at the top layer of the decoder, which is contrary to LCLR and our analysis in Section 3.4 where we argue that the bottom layers of the decoder are more sensitive to the language features.

[5]Note that, those models are trained by adding a source language tag at the encoder and a target language tag at the decoder. In fine-tuning, we keep the original strategy.

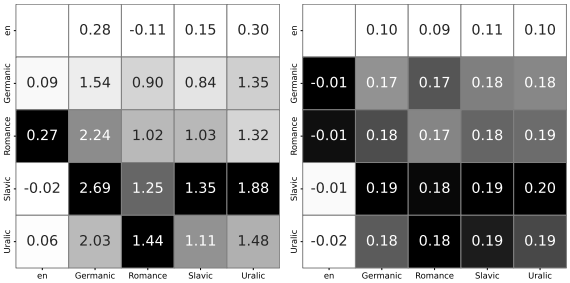
5.2 Results

First of all, Gu et al. (2019); Liu et al. (2021) pointed out that the vanilla Transformer is superior in supervised translation directions, i.e., en<->x, because the model excessively focuses on English, which is the language dominating the training set, to lose its generalization on non-English languages, i.e., the zero-shot translation. Moreover, Huang et al. (2023); Chen et al. (2023b) showed that improving the zero-shot may come at the expense of supervised performance. In this work, our methods significantly improve the zero-shot translation without degrading the supervised performance in both training from scratch and fine-tuning.

Table 1 shows the experimental results of training from scratch. In supervised translations of Europarl-15/TED-19/OPUS-100, LOLE shows divergent results of 0.13/-0.04/-0.22 on en→x and 0.11/0.05/-0.02 on x→en. Similarly, LCLR shows diverse results of -0.05/-0.01/-0.03 and 0.04/-0.01/0.07, respectively. Then, BOTH achieves the results of 0.18/-0.02/0.03 on en→x and 0.12/0.12/-0.03 on x→en. In zero-shot translations, BOTH outperforms VANILLA 1.55/1.33/2.93 for BLEU and 0.58/1.02/3.12 for BERTScore. Our models perform best in zero-shot translations of Europarl-

	BLEU			BERTScore		
Method	en→	→en	zero.*	en→	→en	zero*
M2M-418M	21.88	26.43	14.51	82.52	93.25	79.26
F.T.	26.68	32.95	17.46	84.47	94.30	80.79
LoLE	26.81	33.16	17.52	**84.51**	94.31	80.84
LCLR	26.81	**33.67**†	17.65	84.47	**94.40**	80.88
BOTH	**26.83**	33.63†	**17.68**	84.49	94.38	**80.90**
M2M-1.2B	24.32	28.94	15.95	83.17	93.72	79.75
F.T.	27.71	**34.97**	18.48	84.71	**94.53**	81.14
LoLE	28.29†	34.12	18.67	84.91†	94.48	**81.26**†
LCLR	28.26†	34.54	18.64	84.90†	94.50	81.22
BOTH	**28.37**†	34.59	**18.69**	**84.92**†	94.51	81.23
mBART50	25.28	33.50	6.92	83.93	**94.43**	72.91
F.T.	27.17	33.96	5.58	84.64	94.36	72.96
LoLE	27.19	33.93	7.28†	84.60	94.37	73.86†
LCLR	27.07	34.02	**9.69**†	84.59	94.38	75.31†
BOTH	**27.36**	**34.04**	9.55†	**84.66**	94.36	**75.55**†

Table 2: Averaged scores for experiments of fine-tuning. F.T. means fine-tuning without any trick. * is added to zero. to show it is not a real zero-shot scenario for M2M. The bold number indicates the best result, and the numbers with † are significantly better than F.T.. The off-target ratios are reported in Appendix I.



(a) BLEU (b) SVCCA

Figure 8: Differences between our model and VANILLA. X-axis is the target language family where en is considered solely. Hence, we plot the color ladder by column where the darker the color, the bigger the difference.

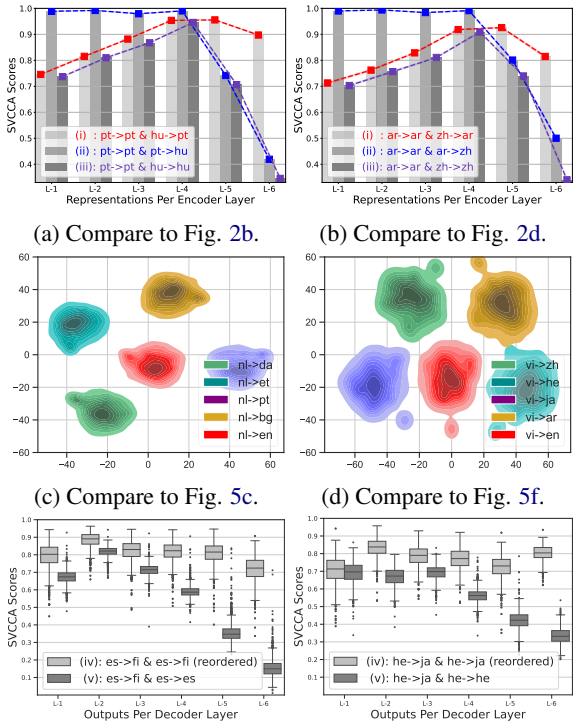
15 and OPUS-100, and the improvements in zero-shot translations are always statistically significant. Note that, although SEMALI achieves the best zero-shot translation performance in TED-19, the supervised performance of SEMALI is significantly degraded compared to VANILLA, which is a common and unresolved problem (Gu et al., 2019; Zhang et al., 2020a; Liu et al., 2021). On the contrary, our methods not only perform competitively with SEMALI in zero-shot translations but also benefit the supervised translation capacity. Moreover, these two proposed methods are orthogonal, which can be proved by assessing LOLE, LCLR and BOTH individually: (1) LOLE achieves gains of 1.53/1.22/2.85 for BLEU and 0.54/0.91/3.34 for BERTScore; (2) LCLR improves 1.06/0.14/0.09 and 0.39/0.13/0.06 scores; (3) The gains of BOTH are always higher than LOLE and LCLR. In addition, we can observe that the improvement of LCLR is limited in TED-19 and OPUS-100, which can be attributed to the diverse languages involving in these two datasets and being easily distinguished by the vanilla decoder. This result also supports that the main challenge of MNMT is the transfer within the encoder. Thus, we can conclude that our methods substantially benefit the zero-shot translation capacity of MNMT models.

Table 2 shows the experimental results of fine-tuning. For M2M-418M, compared with F.T., our methods obtain up to 0.15/0.72/0.22 for BLEU scores and 0.04/0.10/0.11 for BERTScore in en→x, x→en and zero-shot translations, respectively; For M2M-1.2B, the gain is up to 0.66/-0.38/0.21 for BLEU scores and 0.21/-0.02/0.12 for BERTScore; For mBART50, the gain is up to 0.19/0.08/4.11 for BLEU scores and 0.02/0.02/1.69 for BERTScore. Those scores show that the improvement on M2M is marginal compared with training from scratch. This derives M2M is trained by interconnected translation pairs instead of an English-centric dataset, which results in the robust transferability of multilingual representations. However, the degeneration on F.T. of mBART50 shows that fine-tuning drastically influences the zero-shot translation capacity. For instance, the BLUE scores of fr→vi decrease to 11.84 from 20.57 and fr→zh increase to 13.52 from 1.90, but our model obtains 18.47 and 17.19, respectively. Such results and the significant testing indicate again the advantage of our proposed methods in improving multilingual representations for zero-shot translation capacity.

6 Discussion

6.1 Correlation between Representational Disentanglements and Improvements

Table 1 shows the overall results by taking averages across all language pairs, which may overlook pair-specific tendencies. Therefore, we group Europarl-15 by the language families and report the average scores of translating from one language family to another. Figure 8a shows the difference in BLEU scores between our models and VANILLA. As shown in Figure 8b, we also compute the SVCCA scores between the identity of the non-central language and the identity of the central language at the encoder's output and group them in the same manner. Given the similar distribution in Figure



(a) Compare to Fig. 2b. (b) Compare to Fig. 2d.

(c) Compare to Fig. 5c. (d) Compare to Fig. 5f.

(e) Compare to Fig. 6a. (f) Compare to Fig. 6b

Figure 9: Visualizations for the encoder that incorporates LoLE and LCLR, showing improvements compared with VANILLA. Additionally, the model plotted in 9e only incorporates LCLR.

8, we conduct Pearson correlation analysis (Pearson, 1896) of all language pairs instead of language families in Europarl-15, and we compute the coefficients and p-values of Pearson correlation by target languages to maintain fairness. We observe two key points: 1) The coefficient and p-value of en are -0.087 and 0.76, respectively. This result suggests that there is no statistical correlation, which is predictable because x→en is not affected by representational entanglements. 2) The coefficient and p-value of non-central languages are in the ranges of 0.585 to 0.855 and 4e-5 to 0.021, respectively. In more detail, the mean values are 0.770 and 0.002 and the variances are 0.04 and 3e-5, respectively. This analysis proves that the degree of representational disentanglement positively correlates with the improvement of zero-shot translations.

6.2 Analysis of Improved Representation

We measure representation transfer in the model incorporating our proposed methods to verify our findings further. As shown in Figures 9a and 9b, both scenarios exhibit improvements on (i). Meanwhile, Figures 9c and 9d indicate that the entanglement of representations among languages at the encoder is resolved. The evidence suggests that LoLE effectively enhances representation transfer in the encoder. Additionally, (ii) and (iii) in Figures 9a and 9b also achieve higher scores at lower layers of the encoder, which suggests that LoLE indeed makes lower layers of the encoder focus on surface-level information. By contrast, as shown in Figures 9e and 9f, the more stable trend of (iv) in both scenarios suggests that LCLR can improve the capacity of lower layers of the decoder to distinguish languages to improve zero-shot translations. In addition, Appendix F provides the representational analysis for fine-tuning models, which proves that target language features are consistently beneficial in the encoder.

7 Related Works

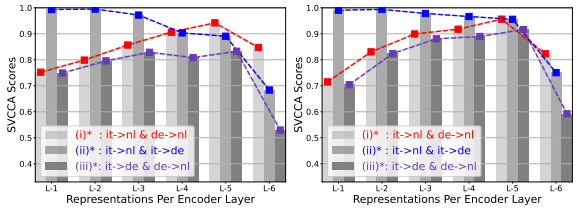
Prior studies on analyzing multilingual representation in Section 2.2 led to several effective methods in MNMT. Some works focused on updating and constraining the encoder to improve multilingual representations, and the findings in discrepancy mentioned in Section 2.2 led to two distinct approaches. First, Pan et al. (2021); Gu and Feng (2022); Gao et al. (2023); Bu et al. (2024) suggested regularizing the encoder for aligning semantic information across different source languages by introducing additional training objectives. Similarly, Pham et al. (2019); Zhu et al. (2020) explicitly modified the output form of the encoder to transfer the representation of the source sentence toward a language-agnostic state. Second, Gu et al. (2019); Liu et al. (2021); Sun et al. (2024) introduced specialized modeling constraints to improve the encoder to transfer source sentence representations to the target language without adding extra parameters, and Zhang et al. (2021); Pires et al. (2023) enhanced the representation of target language information by simply adding language-specific modules. Additionally, Yang et al. (2021); Qu and Watanabe (2022); Bu et al. (2024) focused on improving the target language representation on the decoder side or adding modules specified to the target language to the decoder. Given that the above works can all be encompassed within our analyses, we argue that this work offers insights for future improvements in MNMT. Specifically, enhancing the encoder to transfer source language representations into the target language subspace and align semantic information within those subspaces is the key to improving MNMT.

In addition, a critical factor of this work is the introduction of the identity pair as an analytical tool. Specifically, while identity pairs have been heuristically used in prior works (Tiedemann and Scherrer, 2019; Thompson and Post, 2020; Bu et al., 2024), as an assumed indicator of language-specific representation states, they have not been subject to systematic or quantitative analysis. In contrast, we explicitly define, validate, and utilize identity pairs to probe representational properties in a controlled and measurable way. This not only strengthens the empirical basis of our conclusions but also constitutes an important methodological contribution of this work.

8 Conclusion

We systematically investigated the representational issue of zero-shot translation deficiency in multilingual neural machine translation models. Our analyses show that the encoder transfers translation representations from the source language to the target language, and aligns semantics across different source languages at the target language subspace. We applied engineering practices to verify our findings by proposing two orthogonal methods, which substantially improve the zero-shot translation capacity. Thus, our findings are significant for guiding the improvement of the transferability of multilingual representations.

9 Limitations



(a) Bidirection (b) Unidirection

Figure 10: Illustration of the comparison between the bidirectional and the unidirectional scenarios. 10a has the same model settings with Figure 2, but analyzes the same pairs with 10b.

This work has two limitations. First, the identity pair is a proxy of language representations based on bi-directional training, i.e., each non-central language appears in the encoder and decoder together. Therefore, we designed an additional study to investigate the impact by retraining a model by eliminating nl→en and en→it so that it and nl appear only in the encoder and decoder, respectively, based on the analysis in Section 3.2. Then, we conducted a comparison by taking de as the middle language to perform the role of identity pairs in analysis. As shown in Figure 10, the target language features keep the same trend as shown in Section 3.2 to support our conclusion again, but the influence of source language features increases relatively.

The second limitation is our investigation is based on adding a language tag specified to the target language at the beginning of the source sentence for the encoder. Although this is the de facto MNMT training strategy (Johnson et al., 2017; Aharoni et al., 2019; Arivazhagan et al., 2019a; Gu et al., 2019; Pham et al., 2019; Wu et al., 2021; Yang et al., 2021; Pan et al., 2021; Qu and Watanabe, 2022; Chen et al., 2023a; Gu and Feng, 2022; Gao et al., 2023), the current open-source models (Fan et al., 2020; Tang et al., 2020; Team et al., 2022) are based on another strategy, i.e., adding a source language tag at the encoder side and adding a target language tag at the decoder side. Although, in Section 5, we have shown our proposed methods also benefit models with this strategy, this effectiveness is proved by empirical experiments. Thus, our future work is to investigate the representation transfer of this strategy to guide further improvements.

10 Further Considerations

Ethical Consideration All datasets used in this work are public data, which are proven harmless. Moreover, this work is foundational research and is not tied to particular applications. Thus, there is no ethical risk existed in this work.

Sustainability statement As noted in Appendix C, the GPU used in training individual models is A6000, which has an estimated carbon dioxide emission of approximately 0.13 kg per hour.[6] Specifically, models trained on Europarl-15 and TED-19 required approximately 48 GPU hours, while models trained on OPUS-100 necessitated around 192 GPU hours.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *Preprint*, arXiv:1903.00089.

[6]Measured by https://mlco2.github.io/impact/.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Roee Aharoni, Melvin Johnson, and Wolfgang Macherey. 2019a. The missing ingredient in zero-shot neural machine translation. *Preprint*, arXiv:1903.07091.

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. 2019b. Massively multilingual neural machine translation in the wild: Findings and challenges. *Preprint*, arXiv:1907.05019.

Mengyu Bu, Shuhao Gu, and Yang Feng. 2024. Improving multilingual neural machine translation by utilizing semantic and linguistic features. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10410–10423, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Liang Chen, Shuming Ma, Dongdong Zhang, Furu Wei, and Baobao Chang. 2023a. On the off-target problem of zero-shot multilingual neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9542–9558, Toronto, Canada. Association for Computational Linguistics.

Liang Chen, Shuming Ma, Dongdong Zhang, Furu Wei, and Baobao Chang. 2023b. On the pareto front of multilingual neural machine translation. *Preprint*, arXiv:2304.03216.

Raj Dabre and Sadao Kurohashi. 2019. Mmcr4nlp: Multilingual multiway corpora repository for natural language processing. *Preprint*, arXiv:1710.01025.

Hiroyuki Deguchi, Taro Watanabe, Yusuke Matsui, Masao Utiyama, Hideki Tanaka, and Eiichiro Sumita. 2023. Subset retrieval nearest neighbor machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–189, 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, 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.

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.

Pengzhi Gao, Liwen Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2023. Improving zero-shot multilingual neural machine translation by leveraging cross-lingual consistency regularization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12103–12119, Toronto, Canada. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O.K. Li. 2019. Improved zero-shot neural machine translation via ignoring spurious correlations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1258–1268, Florence, Italy. Association for Computational Linguistics.

Shuhao Gu and Yang Feng. 2022. Improving zero-shot multilingual translation with universal representations and cross-mapping. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6492–6504, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

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.

Yichong Huang, Xiaocheng Feng, Xinwei Geng, Baohang Li, and Bing Qin. 2023. Towards higher Pareto frontier in multilingual machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3802–3818, Toronto, Canada. Association for Computational Linguistics.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. 2022. Understanding dimensional collapse in contrastive self-supervised learning. *Preprint*, arXiv:2110.09348.

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.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *International Conference on Learning Representations*.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

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.

Philipp Koehn et al. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Danni Liu, Jan Niehues, James Cross, Francisco Guzmán, and Xian Li. 2021. Improving zero-shot translation by disentangling positional information. 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 1259–1273, 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. *Preprint*, arXiv:2001.08210.

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.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual 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 244–258, Online. 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.

Karl Pearson. 1896. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 187:253–318.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, and Alexander Waibel. 2019. Improving zero-shot translation with language-independent constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 13–23, Florence, Italy. Association for Computational Linguistics.

Telmo Pires, Robin Schmidt, Yi-Hsiu Liao, and Stephan Peitz. 2023. Learning language-specific layers for multilingual machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14767–14783, Toronto, Canada. 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.

Zhi Qu and Taro Watanabe. 2022. Adapting to non-centered languages for zero-shot multilingual translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5251–5265, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6078–6087, Red Hook, NY, USA. Curran Associates Inc.

Naomi Saphra and Adam Lopez. 2019. Understanding learning dynamics of language models with SVCCA. 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 3257–3267, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

David Stap, Vlad Niculae, and Christof Monz. 2023. Viewing knowledge transfer in multilingual machine translation through a representational lens. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14973–14987, Singapore. Association for Computational Linguistics.

Zengkui Sun, Yijin Liu, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2024. LCS: A language converter strategy for zero-shot neural machine translation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9201–9214, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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.

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. *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, 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.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Yuandong Tian. 2022. Understanding deep contrastive learning via coordinate-wise optimization. In *Advances in Neural Information Processing Systems*.

Jörg Tiedemann and Yves Scherrer. 2019. Measuring semantic abstraction of multilingual NMT with paraphrase recognition and generation tasks. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 35–42, Minneapolis, USA. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

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.

M. P. Wand and M. C. Jones. 1993. Comparison of smoothing parameterizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, 88(422):520–528.

Liwei Wu, Shanbo Cheng, Mingxuan Wang, and Lei Li. 2021. Language tags matter for zero-shot neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3001–3007, Online. Association for Computational Linguistics.

Frank F. Xu, Uri Alon, and Graham Neubig. 2023. Why do nearest neighbor language models work? *Preprint*, arXiv:2301.02828.

Yilin Yang, Akiko Eriguchi, Alexandre Muzio, Prasad Tadepalli, Stefan Lee, and Hany Hassan. 2021. Improving multilingual translation by representation and gradient regularization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7266–7279, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qi Ye, Sachan Devendra, Felix Matthieu, Padmanabhan Sarguna, and Neubig Graham. 2018. When and why are pre-trained word embeddings useful for neural machine translation. In *HLT-NAACL*.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020a. 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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Changfeng Zhu, Heng Yu, Shanbo Cheng, and Weihua Luo. 2020. Language-aware interlingua for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1650–1655, Online. Association for Computational Linguistics.

A Sentence-level SVCCA Score

We use SVCCA (Raghu et al., 2017) to measure representation similarity in MNMT (Kudugunta et al., 2019). We follow the approach of Liu et al. (2021) so that similarity is measured at the sentence level to ensure that each score is computed on equivalent features without the influence of other sentences in the set.

Based on the definition of Section 2.1, we denote hidden representations of a sentence by $\mathbf{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2 ... \boldsymbol{h}_q\}$, where $\mathbf{H} \in \mathbb{R}^{q \times d}$, q equals to the length n or m from either the encoder or decoder, and d is the model dimension. Additionally, the practical length is $n + 1$ when $\mathbf{H}$ is fed into the encoder because the encoder receives the input concatenated by l and $\boldsymbol{x}$[7]. Then, we derive the sentence-level representation $\overline{\boldsymbol{h}}$ using average pooling $\overline{\boldsymbol{h}} = \frac{\sum_{i=1}^{q} \boldsymbol{h}_i}{q}$, which mainly represents the language features and semantics of the source sentence rather than syntactic information because positional information is reduced.

Given $\mathbf{H}^a$ and $\mathbf{H}^b$ derived from two sentences, SVCCA first performs singular value decomposition on their averaged representations to obtain subspace representations $\overline{\boldsymbol{h}}^a \in \mathbb{R}^{d^a}$ and $\overline{\boldsymbol{h}}^b \in \mathbb{R}^{d^b}$, where noise is reduced (Saphra and Lopez, 2019). Then we perform canonical correlation analysis (Hardoon et al., 2004) to determine $\mathbf{W}^a \in \mathbb{R}^{d' \times d^a}$ and $\mathbf{W}^b \in \mathbb{R}^{d' \times d^b}$. Formally, we compute correlation ρ between $\overline{\boldsymbol{h}}^a$ and $\overline{\boldsymbol{h}}^b$ as

$$\rho = \frac{\langle \mathbf{W}^a \overline{\boldsymbol{h}}^a, \mathbf{W}^b \overline{\boldsymbol{h}}^b \rangle}{\|\mathbf{W}^a \overline{\boldsymbol{h}}^a\| \|\mathbf{W}^b \overline{\boldsymbol{h}}^b\|}, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ indicates the inner product. We use ρ to represent the similarity of two sentences. Finally, we compute the set-level score by taking the average scores of all sentences over the set.

B Detailed Information of Datasets

This work involves three datasets, i.e., Europarl-15, TED-19, and OPUS-100, where Europarl-15 and TED-19 are used in preliminary experiments. The training sets of those three datasets have different sizes, but the validation and test sets of a pair generally contain 2,000 translation instances. In preliminary experiments, we measure SVCCA scores in the test sets because those instances are unseen in the training.

Europarl-15 is collected from MMCR4NLP, which has high-quality translation instances and each instance in a language is one-to-one corresponding to other languages, i.e., all language-specific sets have parallel semantics (Koehn et al., 2005; Dabre and Kurohashi, 2019), including 15 European languages from 4 language families. Specifically, Germanic includes `en`, `de`, `nl`, `da`, Romance includes `es`, `pt`, `it`, `ro`, Slavic includes `sl`, `bg`, `pl`, `cs`, and Uralic includes `fi`, `et`, `hu`. The training and validation sets cover 28 supervised translation pairs where English is the central language used to bridge the non-central languages. The test set consists of all language pairs, including 182 zero-shot translation pairs in addition to supervised translation pairs. Finally, each pair in the training set comprises 189,310 instances.

In contrast to Europarl-15, which is the semantically aligned dataset, TED-19 consists of 19 languages, including `en`, `ar`, `he`, `ru`, `ko`, `it`, `ja`, `zh`, `es`, `nl`, `vi`, `tr`, `fr`, `pl`, `ro`, `fa`, `hr`, `cs`, `de`, which belong to various language families without parallel semantics, from TED Talks (Ye et al., 2018). Each translation pair contains 103,093 to 214,111 instances in training, and the training set comprises 6,551,456 instances in total. Because of the unparallel semantics of TED-19, we align `ar`, `he`, `zh`, `hr`, `vi`, `ja` to obtain 967 translation instances for measuring SVCCA scores. In addition, the reason why the number of languages is 19 is that, first, TED Talks have 20 high-resource languages, which are supported in M2M (Fan et al., 2020) and mBART50 (Tang et al., 2020). However, the tokenization of `th` is problematic, resulting in deprecating `th`.

OPUS-100 consists of 95 languages, 188 pairs, and 109.2 million instances in total (Zhang et al., 2020a; Yang et al., 2021), where 90 pairs comprise 1 million instances and 56 pairs have more than 0.1 million instances. Different from Yang et al. (2021), we do not include the zero-shot translation pairs in the validation set to avoid biases when assessing the transferability of multilingual representations.

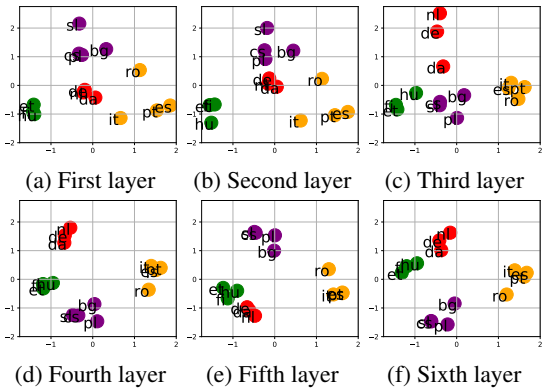
C Detailed Settings of Models

We implement the Transformer (Vaswani et al., 2017) as the backbone model via Fairseq (Ott et al., 2019). For the configuration of models trained on

[7] l plays the role of translation instruction instead of a token belonging to the target language with semantics, thus, this concatenation would not influence the measurement by mixing target language information into the sentence representation within the encoder.

Europarl-15 and TED-19, we follow Kudugunta et al. (2019) to set 6 encoder and decoder layers. Based on the ablation study conducted in the validation set in Europarl-15 and TED-19 shown in Appendix E, we apply LOLE in the fifth encoder layer, set d^e to 128, and set d^h and k of LCLR to 64 and 30, respectively. When we solely apply LCLR, we set the position to the bottom decoder layer based on the findings in Section 3.4. When we integrate both LOLE and LCLR into a model, we relocate LCLR to the second-bottom decoder layer because of the improved language features of the encoder representations. We adopt a shared vocabulary trained by SentencePiece (Kudo and Richardson, 2018) with 50,000 tokens for both the encoder and decoder. The model consists of 4 attention heads, embedding size of 512, inner size of 1024, dropout rate of 0.2, maximum learning rate of 0.0005 with the inverse square root schedule and 4,000 warmup steps, and label smoothing rate of 0.1. We set the batch size to 8,000 tokens per GPU, apply Adam (Kingma and Ba, 2017) as the optimizer, and set temperature sampling with $T = 5$ (Arivazhagan et al., 2019b). We train the model with 60 epochs for Europarl-15 and 30 epochs for TED-15, and finally average the top 5 checkpoints using the loss on the validation set. Compared with the basic configuration, the models trained on OPUS-100 have 8 attention heads, embedding size of 512, inner size of 2048, dropout rate of 0.1, and shared vocabulary size of 64,000. We enlarge d^e to 256 and d^h to 128 for models trained on OPUS-100 and three pre-trained models because they involve more languages. We train the model of OPUS-100 for 400,000 update steps with a batch size of 8,000 tokens per GPU for OPUS-100 and directly use the best checkpoint selected using the loss on the validation set. Furthermore, models with Europarl-15 and TED-19 are trained on 8 NVIDIA V100 GPUs, and models with OPUS-100 are trained on 4 NVIDIA A6000 GPUs by setting *–update-freq* to 2 in Fairseq to simulate 8 GPUs.

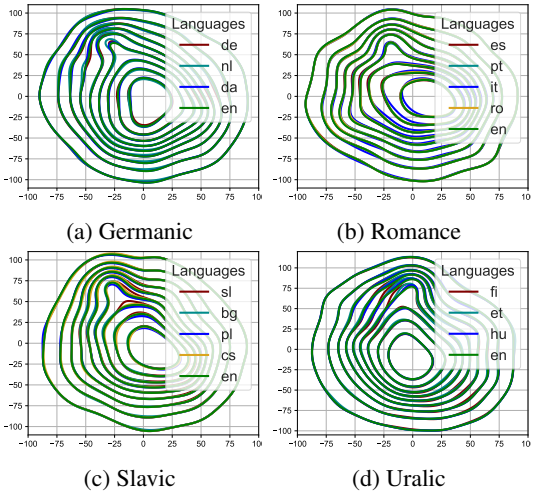
Three open-source models are utilized in fine-tuning experiments. The first is M2M-418M (Fan et al., 2020), trained on standard multilingual translation tasks and supporting translation across 100 languages. It is based on Transformer architecture, configured with 12 encoder and decoder layers, embedding size of 1024, inner size of 4096, and vocabulary size of 128,112, which results in a total of 418 million parameters. The second model, M2M-1.2B (Fan et al., 2020), enlarges the number of layers to 24 and the inner size to 8192 on M2M-418M, and culminates in 1.2 billion parameters. The last model is mBART50 (Tang et al., 2020), trained on monolingual corpora across 50 languages following Lewis et al. (2019); Liu et al. (2020) and preliminarily fine-tuned for MNMT. It shares the same parameter setup as M2M-418M with a vocabulary size of 250,053, which consists of 611 million parameters. We conduct experiments on TED-19 because all covered languages are supported by these models.



(a) First layer (b) Second layer (c) Third layer

(d) Fourth layer (e) Fifth layer (f) Sixth layer

Figure 11: Affinities for en→x at each encoder layer. Language families of Europarl-15 are distinguished by colors: Germanic by red, Romance by yellow, Slavic by purple, and Uralic by green.



(a) Germanic (b) Romance

(c) Slavic (d) Uralic

Figure 12: Visualizations by t-SNE and BiKDE of aligning representations between en→en and x→en of Europarl-15 at the output of the encoder.

D Detailed Introductions of Figure 1

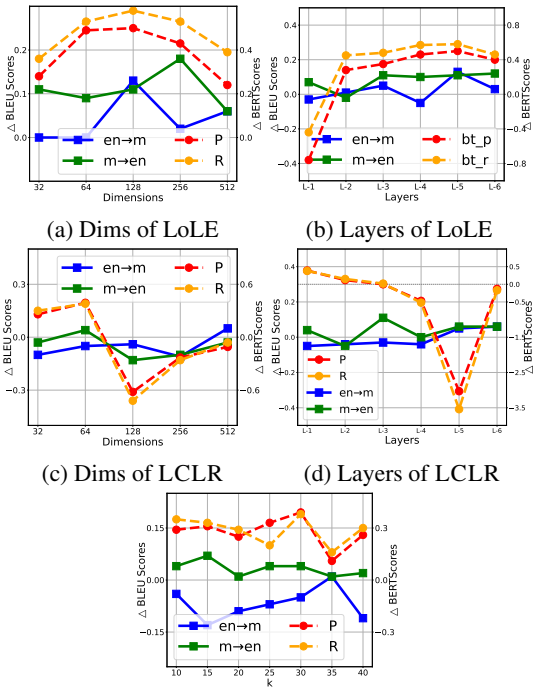
In fact, Figure 1a corresponds to the last subfigure of Figure 11 to show the linguistic affinity between translations from English to other languages, denoted by en→x. Specifically, Fig-

ure 11 shows the layer-wise states of the encoder, and Figure 1a (Figure 11f) demonstrates the state at the output of the encoder. We employ *sklearn.manifold.SpectralEmbedding*, referring to `https://scikit-learn.org`, to visualize the similarities computed by SVCCA (Appendix A) for every layer in the encoder. Then, we can find that representations at all encoder layers have certain clusters influenced by the families of the target languages, and the clusters become more distinct as the depth of the encoder layers increases. This suggests that the transfer of representations to the target language begins as early as the first layer of the encoder, with gradual strengthening through further layers. Meanwhile, this finding, i.e., even the initial encoder layers capture target language features, complements prior works (Kudugunta et al., 2019; Pires et al., 2023).

On the other hand, we follow Pan et al. (2021) and Gao et al. (2023) to measure the alignment of encoder representations between the identity of en and source languages from different families to English using t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008) and bivariate kernel density estimation (BiKDE) (Wand and Jones, 1993). As shown in Figure 12, representations from the four language families are all highly aligned with the identity pair of en→en, where the common feature of those translations is the parallel semantics. Thus, this proves that the encoder semantically aligns different translations. However, the deep discussion should be referred to Section 3.2.

E Selecting Hyper-Parameters

We conduct ablation studies on the validation set of Europarl-15 to select hyper-parameters for LoLE and LCLR, which are used in Section 5.1. Figure 13a shows that LoLE performs optimally with the dimension of 128, which corroborates our hypothesis in Section 4.1. Figure 13b indicates that LoLE performs the best at the fifth layer and degrades significantly at lower layers, which aligns with our assertion in Section 3.2 that lower layers of the encoder are more correlated with the source language, and enhances the language transfer benefits of transferability (Section 4.1). Figure 13c is consistent with the theory of contrastive learning in which full dimensions lead to collapse (Jing et al., 2022). Figure 13d demonstrates that, as the position constructed by LCLR increases, the



(a) Dims of LoLE (b) Layers of LoLE

(c) Dims of LCLR (d) Layers of LCLR

(e) k of LCLR

Figure 13: Illustrations for the ablation study. Δ means the difference between the scores of our methods and the scores of VANILLA. 13a and 13b present variations of LoLE in dimensions and layers, respectively; and 13c, 13d, and 13e present variations of LCLR in dimensions, layers, and k, respectively.

scores decrease, which lends support to our analysis in Sections 3.4 that the instability of decoder representations primarily manifests at lower layers, which also explains the weakness of TLP because improving the capacity of distinguishing languages is redundant for the decoder's top layer. We also conduct an ablation study for hyperparameter k for LCLR with a dimension of 64 at the bottom decoder layer. The results are shown in Figure 13e, with an empirically optimal $k = 30$.

F Analysis of Improved Representation for Fine-tuning Pre-trained Models

Section 6.2 is the representational analysis for models, which are trained from scratch with proposed LOLE and LCLR. We also show the representational analysis for fine-tuning pre-trained models.

Given the positive correlation shown in Section 6.1, we compute SVCCA scores in the same way as done in Section 3.2 and show the results in Table 3. Unlike Section 3.2, we equally consider the encoder and decoder because the encoder is only related to the source language and does not transfer representations to the target language in the training

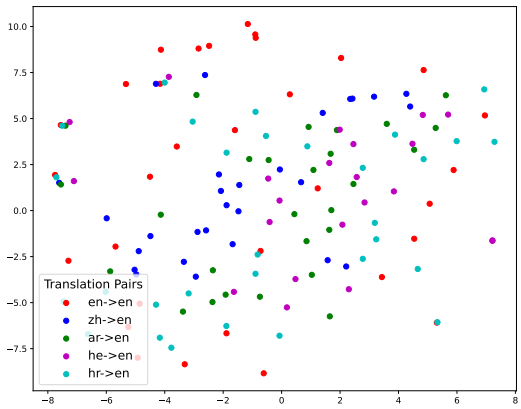
	Pairs	Model	Method	(i)	(ii)	(iii)
Encoder Side	① of zh ② of ar	M2M	F.T.	79.66	100.0	79.66
			LoLE	79.52	98.75	77.76
		mBART	F.T.	63.97	100.0	63.97
			LoLE	63.52	97.90	61.66
	① of he ② of vi	M2M	F.T.	81.50	100.0	81.50
			LoLE	80.54	98.27	79.88
		mBART	F.T.	69.17	100.0	69.17
			LoLE	70.46	97.61	67.04
Decoder Side	① of ja ② of he	M2M	F.T.	99.80	92.01	92.65
			LoLE	99.73	89.81	90.66
		mBART	F.T.	98.53	90.76	89.62
			LoLE	98.64	90.07	88.49

Table 3: SVCCA scores. Each score times 100 for a clear illustration. (i) compares the identity of ① and ①→②, (ii) compares the identity of ① and ②→①, and (iii) compares identities of ① and ②. Encoder Side means computing the output of the encoder, and Decoder Side means computing the output of the 1st layer of the decoder.

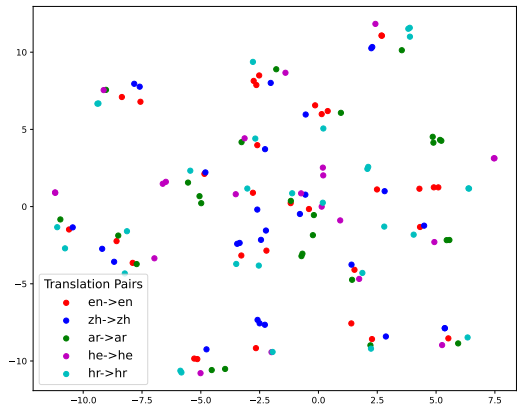
strategy of M2M (Fan et al., 2020) and mBART50 (Tang et al., 2020). Additionally, the different training strategy is the primary reason that F.T. shows the same scores in (i) and (iii) and keeps 100.0 in (ii). Alternatively, although the scores of (i), which reflect target language features, decrease in our methods, the scores of (ii) and (iii) also decrease. As a result, the differences between the scores of (i), (ii), and (iii) increase, that is, the relative importance of target language features increases. This result proves our statements in Sections 3.3 and 6.1 again that target language features are consistently beneficial in the encoder. On the other hand, the decoder side shows the same tendency as the encoder side. This fits our motivation in Section 4.2 to further improve the discriminating ability of lower layers of the decoder, although the training strategy of M2M and mBART50 has already provided a high capacity in discrimination for the decoder.

G Token-level Alignments in Other Cases

First of all, the English sentence for semantic analysis in Figures 4 and 14 is: By the end of this year, there will be nearly a billion people on this planet that actively use social networking sites. Compared with the discussion in Section 3.2, token-level representations are not aligned at the embedding layer, and are relatively aligned in the case of using the identity pairs, where the degree of divergence is substantially higher than the case of Figure 4.



(a) Semantic alignments on embeddings



(b) Semantic alignments on identities

Figure 14: Illustration of the token-level alignment corresponding to Figure 1b. Representations shown in 14a are collected at the embedding layer, whose overall variance is 1.45. Representations shown in 14b are collected from identities, whose overall variance is 0.13.

H Evaluation Metrics Selection

In this work, we select two main automatic evaluation metrics and a secondary statistic measurement. The first one is SacreBLEU (Post, 2018) which is an implementation of BLEU (Papineni et al., 2002). This is the most popular and common metric used in evaluating the alignment between inferences and references at the word level. In order to counter the insufficiency of SacreBLEU, we also select BERTScore (Zhang et al., 2020b), which is a representational metric to evaluate the semantic similarity between inferences and references. Furthermore, to show whether the improvements brought by proposed methods are significant, we also conduct the statistical significance testing (Koehn, 2004) using paired bootstrap resampling with 1,000 iterations and 0.5 resampling ratios, consequently, the case of $p < 0.05$ means that the difference is significant.

Additionally, we follow prior works (Yang et al.,

	Europarl-15		TED-19		OPUS-100	
Method	zero.(↑)	off.(↓)	zero.(↑)	off.(↓)	zero.(↑)	off.(↓)
VANILLA	24.65	1.34	11.98	4.08	5.04	70.41
DISPOS	25.89	0.84	12.80	3.82	5.58	61.65
TLP	24.96	1.22	12.74	3.71	4.60	83.29
SEMALI	25.25	0.99	13.45	3.62	6.42	58.25
LOLE	26.09	0.71	13.20	3.69	7.92	50.05
LCLR	25.71	0.79	12.12	3.86	5.11	68.53
BOTH	26.20	0.74	13.31	3.69	7.97	55.06

Table 4: Off-target ratio corresponding to experimental results in Table 1. zero. indicates the BLEU scores of zero-shot translations. off. indicates the off-target ratio counted by all zero-shot translation pairs.

Model	Metric	PRE.	F.T.	LOLE	LCLR	BOTH
M2M-418M	zero.(↑)	14.51	17.46	17.52	17.65	17.68
	off.(↓)	3.66	3.34	3.32	3.24	3.33
M2M-418M	zero.(↑)	15.95	18.48	18.67	18.64	18.69
	off.(↓)	3.50	3.15	3.16	3.16	3.14
mBART50	zero.(↑)	6.92	5.58	7.28	9.69	9.55
	off.(↓)	43.56	65.26	40.76	38.24	35.28

Table 5: Off-target ratio corresponding to experimental results in Table 2. Abbreviations follow Table 2 and PRE. refers to the model without any fine-tuning. In addition, compared to Table 4, we switched the horizontal and vertical axes, because there is only one dataset, TED-19, used in fine-tuning experiments.

2021; Chen et al., 2023a) to report the off-target ratio, which is measured by *fasttext-langdetect*[8]. The off-target translation refers to a sentence translated to an incorrect target language rather than the target language we expected. However, the off-target ratio is not reliable, because the popular tools used in measuring off-target ratios are based on word level and lack support in low-resource languages. Furthermore, the score of SacreBLEU can directly show the problem of off-target, because the evaluation process of SacreBLEU tends to give a great penalty on an inference, which has a different writing script from the expected target language. Therefore, we only report it as a secondary metric in Appendix I.

I Off-Target Ratio of Results

Tables 4 and 5 show the measurement of the off-target ratio, which are the supplement of Tables 1 and 2. We can observe that the off-target ratio is always inversely proportional to BLEU scores, aligning with our discussion in Appendix H. Additionally, there are two points worth noting: (1) In Table 4, the off-target ratio in OPUS-100 is generally higher. This is not an outlier because resulting in a strong zero-shot translation capability in OPUS-100 is particularly challenging due to the large number of languages involved and the limited corpus for individual languages (Zhang et al., 2020a; Yang et al., 2021). (2) In Table 5, the off-target ratio counted from mBART50 is higher than other cases. This abnormal value has been discussed in Section 5.2, that is, the zero-shot ability of mBART50 is weaker than M2M models, and then, the fine-tuning dramatically changes the behavior of the model.

[8] https://pypi.org/project/fasttext-langdetect

Decoding Machine Translationese in English-Chinese News: LLMs vs. NMTs

Delu Kong[1,2] **and Lieve Macken**[2]

[1]School of Foreign Studies, Tongji University, Shanghai, 200092, China
[2]Language and Translation Technology Team, Ghent University, Ghent, 9000, Belgium
Correspondence: kongdelu2009@hotmail.com

Abstract

This study explores Machine Translationese (MTese) — the linguistic peculiarities of machine translation outputs — focusing on the under-researched English-to-Chinese language pair in news texts. We construct a large dataset consisting of 4 sub-corpora and employ a comprehensive five-layer feature set. Then, a chi-square ranking algorithm is applied for feature selection in both classification and clustering tasks. Our findings confirm the presence of MTese in both Neural Machine Translation systems (NMTs) and Large Language Models (LLMs). Original Chinese texts are nearly perfectly distinguishable from both LLM and NMT outputs. Notable linguistic patterns in MT outputs are shorter sentence lengths and increased use of adversative conjunctions. Comparing LLMs and NMTs, we achieve approximately 70% classification accuracy, with LLMs exhibiting greater lexical diversity and NMTs using more brackets. Additionally, translation-specific LLMs show lower lexical diversity but higher usage of causal conjunctions compared to generic LLMs. Lastly, we find no significant differences between LLMs developed by Chinese firms and their foreign counterparts.

1 Introduction

A recent, striking report – with an arguably sensational title – proclaims a groundbreaking milestone for LLMs in the field of machine translation (MT): *"Machine Translation is Almost a Solved Problem"*[1]. Although the article's perspective is primarily forward-looking, with a clear acknowledgment on the enduring value of human translation, its message to the public, as the title suggests, is rather obvious: With the help of LLMs, MT currently appears to be nearing perfection. But is it?

Multiple studies showed that LLMs have revolutionized the way we approach language translation, reaching to an unprecedented level of accuracy, contextual understanding, and fluency (Jiao et al., 2023; Peng et al., 2023; Wang et al., 2023; Enis and Hopkins, 2024). It even outperformed some specialized NMT systems under a fine-grained evaluation setting (Manakhimova et al., 2023). Further compelling evidence of the benefits of LLMs is reflected in the WMT24 finding summary (Kocmi et al., 2024), which clearly demonstrate their dominance in the competition. Most of the top-performing systems were LLM-based, with standout models, like Claude-3.5-sonnet, achieving leading positions across multiple language pairs.

Despite their advantages, several limitations persist. Notably, LLMs often face challenges with low-resource languages (Enis and Hopkins, 2024), in explaining, practicing and translating sophisticated concepts (Qian and Kong, 2024), and addressing gender bias issues associated with vocabulary options (Stafanovičs et al., 2020). An area that remains largely underexplored is the influence of machine translationese. While MTese has been demonstrated in NMT output (Vanmassenhove et al., 2019, 2021), it has not yet been fully investigated in the context of LLMs.

MTese can subtly influence the readability, naturalness, and even credibility of news articles, potentially shaping public perceptions. Studying MTese is critical from several perspectives. In education, MT has been widely utilized in second language acquisition, with MT output sometimes even regarded by students as "expert" (Rowe, 2022). However, MTese may potentially impact the authenticity of learning materials, raising concerns about its influence on learners' lack of exposure to genuine linguistic patterns. In language evolution, it provides insights into how machine-mediated commu-

[1]https://www.economist.com/science-and-technology/2024/12/11/machine-translation-is-almost-a-solved-problem

Proceedings of Machine Translation Summit XX Volume 1, pages 99–112, June 23-27, 2025

nication might drive changes in linguistic norms, and whether MTese would also plays a part in the course, such as influencing language complexity (Cristea and Nisioi, 2024). In literary translation, MTese poses a potential hindrance to the creativity and linguistic richness of literary translation, continuously challenging the long-debated concept of "human parity" (Poibeau, 2022) from a stylistic perspective.

Against this background, we have chosen MTese as the focus of this study, specifically exploring its manifestations and differences in NMTs and LLMs for the linguistically distant language pair of English to Chinese (E2C).

2 Related work

First introduced by Gellerstam (1986), the concept of translationese describes the systematic influence of a source language on the target language during translation. When applied to MT, Daems et al. (2017) emphasize the pivotal role of MTese in shaping the characteristics of post-edited texts, analyzing 55 linguistic features ranging from POS tags to dependency parsing.

Expanding on this, Toral et al. (2018) explore lexical density and diversity, revealing that post-edited literary MTs tend to be more simplified, normalized, and influenced by the source text compared to human translations (HTs), where MT outputs exhibit lower lexical density than HTs, with the neural system showing even lower density than those from the statistical system. However, Castilho et al. (2019) report contrasting findings from a different genre. For news texts, MTs show slightly higher lexical density and richness than HTs, whereas for literary texts, MTs demonstrate slightly lower lexical density but comparable lexical richness to HTs.

In a similar vein, Loock (2020) investigate MTese by analyzing linguistic deviations in English-to-French MT texts compared to original, untranslated texts, providing a broader perspective on the systematic over-representation of linguistic features and their implications for translator training and post-editing practices. De Clercq et al. (2021), working on the same language pair, used 22 linguistic features to distinguish between the original and MTed French. They showed that average sentence length and four features related to formulaicity could discriminate between original and MTed French.

However, for linguistically distant language pairs like E2C, research on MTese remains relatively sparse. Jiang and Niu (2022) examine a corpus of English translations of modern Chinese literary texts, including texts translated by NMT and humans. They confirm the presence of translationese in both human and machine translations compared to original texts in some coherence metrics. A recent study by Niu and Jiang (2024) revealed that simplification is a notable characteristic of NMT texts across genres in the E2C direction, such as a loss of lexical complexity.

A general conclusion drawn from the works above is that translations produced by MT engines consistently exhibit a loss of lexical and syntactic richness (Vanmassenhove et al., 2019; Castilho and Resende, 2022). Studies tend to apply fine-grained linguistic features to reveal consistent distributional patterns. Similar phenomena are also observed across various language pairs.

Despite these findings, it remains unclear whether LLMs exhibit distinct features of MTese or whether their outputs can be reliably differentiated from those of NMT engines, particularly in general text types as news discourse. Therefore, our study addresses this gap by focusing on the distant language pair of E2C, emphasizing general news texts, and constructing a larger and more comprehensive dataset and feature set for analysis. We adopt the study design of Loock (2020) and De Clercq et al. (2021), and compare MT [2] outputs generated by different systems with original texts.

We address the following research questions:

- RQ1: Does MTese exist in E2C MTed news texts (NMTs and LLMs)? If so, which linguistic features contribute most to this distinction?
- RQ2: How do LLMs differ from NMTs in their manifestation of MTese across linguistic features?
- RQ3: Do translation-specific and generic LLMs differ from each other? Additionally, how do LLMs developed by Chinese companies compare with those developed by foreign companies in this regard?

3 Methodology

3.1 Dataset

The dataset used in this study encompasses four corpora (detailed information is shown in Table 1), representing original Chinese texts and E2C MTed news texts. The original Chinese news cor-

[2] In this paper, the term MT is used as a superordinate term for both NMT and LLM translation, while NMT and LLM could also be treated as distinct categories.

pus is sourced from two authoritative outlets, *People's Daily*[3] (人民日报) and *Xinhua News*[4](新华网), while the original English news corpus includes articles from reputable platforms like *The Economist*[5] and *The Guardian*[6]. The selected corpora consist exclusively of news texts published after 2022 to avoid the potential influence of outdated news that may have been incorporated into the training data of LLMs. The sample length within each original corpus is maintained around 900 words in average for both Chinese and English.

All texts underwent careful preprocessing, including cleaning, denoising, part-of-speech (PoS) tagging, and dependency (Dep) tagging. Chinese is a language without explicit word boundaries, which requires word segmentation in advance. To achieve state-of-the-art performance, we utilized the Language Technology Platform (LTP)[7], a comprehensive natural language processing toolkit (Che et al., 2021). We used LTP's advanced deep learning models (Base2) to perform word segmentation, PoS tagging, and syntactic analysis. Its reported performances reach 99.18%, 98.69%, and 90.19% for segmentation, PoS tagging, and Dependency parsing, respectively.[8]

The MT data are generated by translating OEN articles into Chinese using each engine on a one-by-one basis, processing each text individually with each engine, thereby minimizing potential interference from varying text topics. As a result, each MT engine produces approximately 200 translations[9]. The dataset includes five NMT engines, comprising three international systems (Google Translate, DeepL, and Microsoft Translator) and two Chinese-developed systems (Baidu Translate and Youdao Translate). Additionally, six LLMs are incorporated, including models developed by Chinese firms (Kimi and ChatGLM), one tailored for MT-specific applications (TowerInstruct), and leading models such as ChatGPT, Claude, and Gemini. All these systems represent SOTA engines on the LLM arena[10] at the time of the experiment (October to December 2024).

The MT process for LLMs involves prompt engineering, with prompt design following Andrew Ng's course guidelines[11] and the CRISPE framework[12]. This approach resulted in a standardized and structured user instruction, which was consistently applied across all engines during the translation process. The full prompt is provided in Appendix A for reference.

To address the potential issue of the OCN corpus having insufficient coverage and variability when limited to the same sample size (200) as other subcorpora, we adopted the corpus structure outlined in De Clercq et al. (2021) and increased the size of the OCN by incorporating more original Chinese news to 2,000 texts in total. This expansion ensures a comparable dataset size between OCN and MTs. Also, we did not impose strict limitations on specific topics within the news genre. Constraining the dataset to a particular domain could lead to data scarcity, as certain topics may not be consistently available over the given period. To maintain balance and comparability, we ensured that all selected news articles were consistent in terms of lexical length and time period.

3.2 Feature set

This section outlines the feature set used in this study. The primary aim of this study is to quantitatively compare different linguistic features across original and MT texts.

Based on the principles of constructing feature sets for translationese studies(Volansky et al., 2013) and referring to previous research (See Huang and Liu, 2009; Lynch and Vogel, 2018; Toral, 2019; De Clercq et al., 2021), the following section presents the feature set used in this study. A brief feature summary can be viewed in Table 4 (in Appendix B). All together, we have employed 236 features in this study. It should be noted that all features are represented as ratios or weighted measures to mitigate the influence of sample size differences and ensure comparability across texts.

Lexical features General lexical features involves common lexical characteristics such as Type-Token Ratio (TTR). The purpose of these features is

[3] http://www.people.com.cn/

[4] http://www.xinhuanet.com/

[5] http://www.economist.com

[6] https://www.theguardian.com

[7] https://github.com/HIT-SCIR/ltp

[8] https://github.com/HIT-SCIR/ltp/blob/main/README.md

[9] It should be noted that several LLMs did not translate all 200 English news (as in Table 1). Some news articles remain untranslated due to "unsafe content" warnings, primarily involving topics related to war or politics. Even though we stated clearly in our prompt that they do not contain any unsafe content. (See Appendix A)

[10] https://lmarena.ai/

[11] https://learn.deeplearning.ai/courses/chatgpt-prompt-eng

[12] https://github.com/mattnigh/ChatGPT3-Free-Prompt-List

Corpus	Type	Abbr.	Engine	Acquisition	Texts	Token	Type
Original	Orig. Chi. News	OCN	-	WebCrawl	2,000	1,685,526	67,082
	Orig. Eng. News	OEN	-	WebCrawl	200	190,572	20,459
NMTs	GoogleTranslate	NGT	-	API	200	171,448	15,288
	DeepL	NDL	-	API	200	177,866	15,104
	MicrosoftTranslator	NMS	-	API	200	172,026	14,586
	BaiduTranslate*	NBD	-	API	200	174,962	14,323
	YoudaoTranslate*	NYD	-	API	200	174,504	16,515
LLMs	ChatGPT	LCG	GPT-4o	Web	200	159,015	14,947
	Claude	LCL	3.5-sonnet	Web	200	170,236	15,123
	Gemini	LGM	1.5-flash	API	189	166,631	14,777
	Kimi*	LKM	moonshot	API	185	145,976	13,225
	ChatGLM*	LGL	GLM-4-plus	API	178	145,700	14,907
	TowerInstruct†	LTO	7B-v0.2	OpenSource	200	175,681	15,398
MTs	In total	MTs	NMTs + LLMs	-	2,152	1,834,045	31,863

Table 1: Overview of the datasets used in this study. Engines marked with an asterisk (*) are primarily trained and tested in mainland China, while the engine in the LLMs marked with a dagger (†) represents an translation-specific model.

to provide an overview of lexical usage in terms of diversity, complexity, and richness. PosTag-based features are derived from the annotation tag set of the LTP platform[13]. For instance, the proportions of nouns or verbs.

Syntactical features General syntactical features focus on capturing broad syntactic patterns and sentence structures in the text, such as average words per sentence. DepTag-based features are built upon the dependency tag set of the LTP platform[14], which identifies the dependency role of each word in a sentence, such as the ratio of verb-object (VOB) and attributive modifiers (ATT).

Readability features Nine readability features proposed by Lei et al. (2024)[15] are included, evaluating lexical, syntactic, and semantic variability to assess the text's difficulty and comprehensibility for the target audience. Complementing these are 4 concreteness features measuring lexical concreteness based on Xu and Li (2020).

Translatibility features The translatability features evaluate linguistic coherence and translation quality between English and Chinese texts through five features: completeness, foreignness, code-switching, abbreviation, and untranslated. Core features such as completeness check for untranslated English sentences longer than three words. Foreignness calculates the ratio of English to Chinese characters.

N-POS-gram features To ensure that the feature set remained content-independent and focused on grammatical patterns rather than topical content, we employed N-PoS-grams (with N ranging from 1 to 3) instead of lexical n-grams. These features capture sequences of part-of-speech tags to highlight grammatical collocations across texts. To refine the selection and reduce the influence of highly frequent but less informative elements (e.g., function words), we used the The Lancaster Corpus of Mandarin Chinese (LCMC[16]) as a reference corpus for comparison. For consistency, the LCMC corpus was re-tagged using the same LTP tools to ensure an aligned PoS tag set.

3.3 Algorithms

3.3.1 Feature selection

To reduce complexity, minimize feature noise, and improve experimental efficiency, a chi-square (χ^2) ranking-based feature selection method is employed in both classification and clustering experiments. Features are ranked based on χ^2 values, and the top-k features are selected, where $k = 30$. If the total number of features in a specific category is less than 30, all available features are retained. This feature selection process mainly reduces the lexical features, N-POS-gram features, and the combined "all features" set in classifying and clustering.

[13] https://ltp.ai/docs/appendix.html#id2
[14] https://ltp.ai/docs/appendix.html#id5
[15] https://github.com/leileibama/AlphaReadabilityChinese
[16] https://www.lancaster.ac.uk/fass/projects/corpus/LCMC/

3.3.2 Classification experiment

The classification experiment is structured based on the hierarchical levels of feature sets. First, classification is conducted using individual feature level, followed by classification using all feature levels. The experiments are organized into the following comparison groups: (1) OCN vs. MTs, where MTs include both NMTs and LLMs subgroups; (2) OCN vs. NMTs and OCN vs. LLMs; (3) LLMs vs. NMTs; and (4) intra-group classifications within the NMTs and LLMs.

Five classifiers, Naïve Bayes, Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest, are employed, and the average classification performance is calculated across these classifiers to provide a balanced result. SVM uses a linear kernel, while the other classifiers follow default settings. Referring to Rahman et al. (2024), the performance of the ensemble classifier is evaluated using Accuracy (ACC) and F1 scores, computed as follows:

$$\text{ACC}_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N}\text{ACC}_i, \quad F1_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N}F1_i$$

Where ACC_i and $F1_i$ are the Accuracy and F1 scores for the i-th classifier, and N is the total number of classifiers. All classification tasks, except intra-group classifications within the NMTs and LLMs groups, are binary classification tasks.

3.3.3 Clustering experiment

The clustering experiment employs the k-means algorithm to cluster the data into three categories: OCN, LLMs, and NMTs. The number of clusters (k) is set to 3, and the Euclidean distance is used to measure the similarity between data points. The top-k significant features selected in prior analysis, are utilized as the feature set for clustering.

To evaluate the clustering performance, the Adjusted Rand Index (ARI) is used as the primary metric. ARI measures the similarity between the clustering results and the ground truth labels, adjusted for chance, providing an objective assessment of clustering quality (Warrens and van der Hoef, 2022). Additionally, Python's Plotly library is employed to generate interactive clustering plots. This approach complements the classification methods, offering an intuitive visualization of the relationships between categories.

4 Results

4.1 Classification

Table 2 presents the results across feature levels and groups. We observe two tendencies:

(1) OCN versus other groups consistently achieves the highest accuracy (around 99% in all feature categories), regardless of whether MT is combined into one group or separated into NMTs and LLMs. Then the performance declines for LLMs-NMTs comparisons (around 70% ACC). The lowest accuracy is observed in intra-group comparisons for both LLMs and NMTs, dropping to below 50%, lower than random distribution baseline.

(2) As for feature categories, lexical, syntactical and N-POS-gram features make the most significant contributions to the classification performance. Take OCN-MTs as an example, the three sets of features all reach to more than 97%. In contrast, readability and translatability features show limited contributions, with accuracies of 86.87% and 70.66% respectively. Combined all features yields the highest overall accuracy (98.92%), which demonstrates the complementary effects of integrating multiple feature categories.

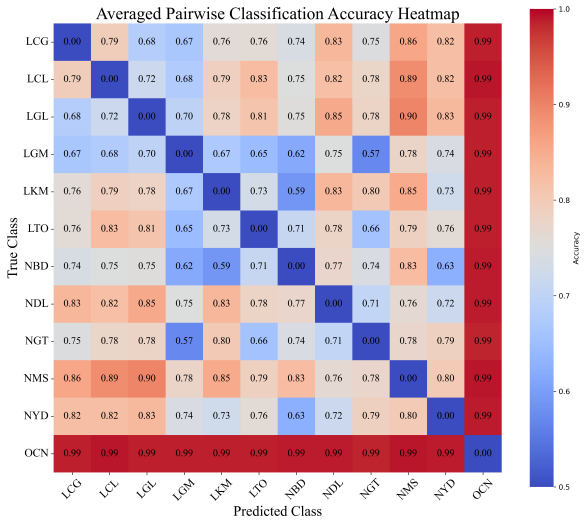


Figure 1: Pair-wise comparison of different MT engines based on 5 averaged classifiers and top 30 salient features

Figure 1 presents a pairwise classification accuracy heatmap to provide a visualized plot and a fine-grained classifying result, using the top-30 salient features from all feature levels. The classification performance is averaged across the five above-mentioned classifiers. For each pair of classes, ACC values are computed and aggregated

Feature Level	Metrics	OCN-MTs	OCN-NMTs	OCN-LLMs	LLMs-NMTs	NMTs (Intra-group)	LLMs (Intra-group)
Lexical	ACC	97.66%	97.77%	97.33%	61.91%	30.04%	30.26%
	F1	0.9766	0.9751	0.9714	0.6144	0.2893	0.2966
	C/T	4054/4152	2932/3000	3067/3152	1332/2152	300/1000	348/1152
Syntactical	ACC	98.46%	98.23%	98.23%	60.53%	36.72%	31.89%
	F1	0.9846	0.9801	0.9809	0.5736	0.3574	0.3022
	C/T	4087/4152	2946/3000	3096/3152	1302/2152	367/1000	367/1152
Readibility	ACC	86.87%	87.61%	85.41%	55.31%	19.44%	25.49%
	F1	0.8683	0.8603	0.8426	0.5452	0.1896	0.2437
	C/T	3607/4152	2628/3000	2691/3152	1190/2152	194/1000	293/1152
Translatibility	ACC	70.66%	78.73%	78.32%	58.26%	26.36%	20.73%
	F1	0.6265	0.6543	0.6618	0.4985	0.2146	0.1515
	C/T	2933/4152	2362/3000	2468/3152	1253/2152	263/1000	238/1152
N-POS-gram	ACC	97.27%	96.51%	96.73%	62.49%	24.26%	24.32%
	F1	0.9603	0.9469	0.9500	0.5287	0.2031	0.1301
	C/T	3987/4152	2865/3000	3008/3152	1322/2152	232/1000	212/1152
All Features	ACC	98.92%	98.84%	98.69%	69.38%	42.10%	35.59%
	F1	0.9891	0.9870	0.9859	0.6902	0.4136	0.3475
	C/T	4106/4152	2965/3000	3110/3152	1492/2152	421/1000	409/1152

Table 2: Performance metrics across feature levels and groups. ACC refers to Accuracy, F1 is a balanced score of precision and recall, while C/T stands for Correctly classified sample / Total samples.

to produce the final heatmap. It reveals three main results:

(1) The deep red along the OCN comparisons highlights its distinctiveness, achieving near-perfect classification accuracy (avg. ACC is close to 0.98) against both LLMs and NMTs.

(2) A similar trend is found in both LLMs and NMTs intra-groups, reflected in the predominantly blue and light orange colours, which stand for 0.6 - 0.8 ACC according to the heatmap legend. For LLMs, ACC ranges from 0.65 to 0.83 (avg. 0.73). Similarly, NMTs exhibit ACC ranging from 0.63 to 0.83 (avg. 0.75)[17].

(3) ACC between LLMs and NMTs is slightly higher (avg. 0.77), indicating more distinct differences between these two groups. The lowest ACC is found between NGT and LGM (avg. 0.57), perhaps due to similar training data since they are both developed by Google[18]. And the highest is found between NMS and LGL (0.90). Notably, NMS stands out with a slightly higher classification ACC against other LLMs (around 0.84).

4.2 Clustering

Figure 2 portraits the clustering results, with an ARI value of 0.64, showing a clear separation of the OCN group (green cluster) on the right, while the left side contains partially overlapping red and blue clusters, primarily NMTs and LLMs samples. This indicates that, if divided into only two clusters, the distinction between OCN and MTs (NMTs and LLMs combined) is more evident. However, within the MT group, there is significant overlap between NMTs and LLMs. The clustering result echoes with the findings in the previous classification experiments.

5 Discussion

5.1 Original Chinese vs. MTs

To answer RQ1, the analysis of Figure 1 and Table 2 reveals significant differences between OCN and MTs (including both NMTs and LLMs) under a sample size of approximately 2000 texts. This indicates that, despite prompt engineering, the trans-

[17]Compared with Table. 2, the higher scores in pairwise models are due to binary classification tasks, which reduces task complexity and better captures discriminative features, whereas multi-class task involves increased feature overlap and requires generalization across all categories.

[18]There exist a possibility of incorporating LLM technology into commercial NMTs, but the specific technical details remain unknown when the company does not disclose further information. So we only select NMT engines as "pure" as possible in our study. For example, the API we use for NGT is v2 (See https://cloud.google.com/translate/docs/editions), which is an NMT engine, rather than v3, which includes LLM.

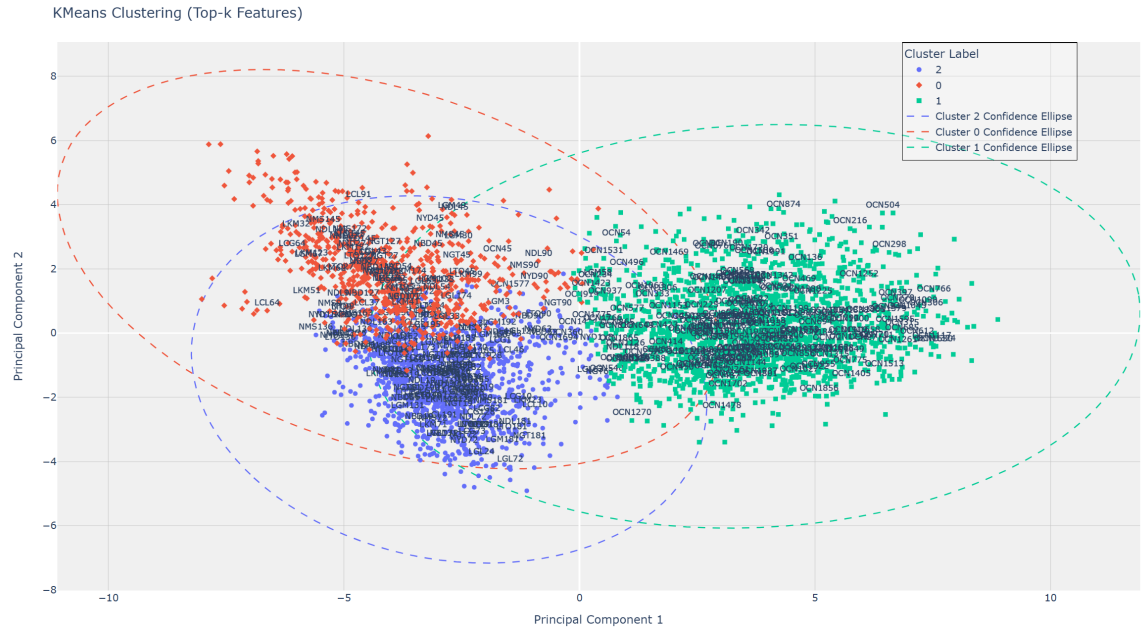


Figure 2: K-means clustering using the top-47 shared features, obtained after deduplication of the top-30 salient features across OCN, NMTs, and LLMs pair-wise comparisons. ARI score: 0.6355.

lations produced by LLMs still exhibit substantial differences from original texts.

Furthermore, the classification results for OCN-NMTs and OCN-LLMs both achieved around 99% ACC. Consequently, there is insufficient evidence to argue that LLMs outperform NMTs in terms of MTese reduction (if they do, then ACC score of OCN-LLMs should be smaller than OCN-NMTs). In the E2C news translation task, while LLMs are often praised for their human-like language abilities in translation (He et al., 2024), their outputs still diverge from authentic Chinese texts. The following analysis explore two prominent features in more detail.

As can be seen in Table 5 of Appendix C, in the OCN-MT group, all top 3 features are related to sentence length, either measured as characters, words or nodes. Therefore, we select the first feature for further elaboration. As shown in Figure 3, there is a significant difference[19] in number of characters per sentence between OCN and MTs, with a Kruskal-Wallis F score reaching to more than 1255. Overall, OCN texts contain more characters per sentence, with a median of 50, compared to less than 40 in MT texts. This echoes with Jiang and Niu (2022) in indicating a preference for shorter

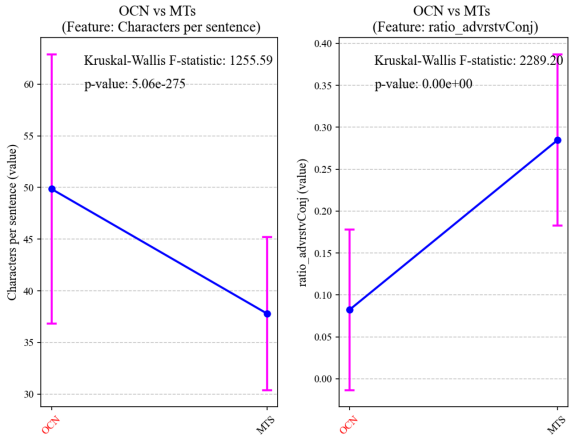


Figure 3: Linguistic differences between OCN and MTs. The left panel compares characters per sentence, while the right panel examines adversative conjunction ratio.

sentences in MT outputs. Additionally, the standardized deviation in OCN spans a wider range (approximately 37 to 63), while MT texts have a narrower range (around 30 to 45), thus there is a greater sentence length variability in original Chinese, yet a more constrained pattern in MTs.

Another interesting finding is that adversative conjunctions are used significantly more frequently in MT texts compared to OCN. In this study, adversative conjunctions are defined as linguistic elements that convey contrastive meanings, such as “但是” (but), “但” (yet), “然而” (however), “可是” (nevertheless), and they could be used inter-

[19]To determine significant differences, we first conduct a normality test on the data. If the data met the normality assumption, we apply ANOVA; otherwise, we use the non-parametric Kruskal-Wallis test.

changeably. As the right side of Figure 3 shows, MTs employ adversative conjunctions more than twice as often as OCN. This phenomenon may be attributed to two factors. First, the difference likely reflects source language interference. OCN articles tend to use fewer adversative conjunctions, while OEN articles, which serve as the source for MTs, employ them more frequently. Second, in handling adversative conjunctions, OCN relies on syntactic transformations or rhetorical devices to reduce their usage. In contrast, both NMTs and LLMs typically employ literal translations of these conjunctions, lacking the ability to restructure sentences to balance their occurrence.

5.2 LLMs vs. NMTs

In terms of RQ2, we address this question in the following two aspects.

Translations generated by LLMs and NMTs share certain linguistics characteristics as classification accuracy is only about 70%. Clustering experiments also reveal that the two systems overlap. This could be attributed to three reasons. (1) Both systems translate from the same OEN articles, meaning that their content and style are inherently influenced by the original text. Thus the differences are largely constrained by the limitations of the source text. (2) Although LLMs utilize extraordinarily large pre-trained data and updated algorithms (Brown et al., 2020), their underlying architecture is based on the Transformer model (Devlin et al., 2019), which was originally applied in NMT systems. (3) It is possible that LLMs utilize training data from NMT systems developed by the same company (e.g. NGT and LGM). Alternatively, NMT systems may have already integrated certain technologies and algorithms from LLMs. All these factors further blur the lines between the two.

Translations generated by LLMs and NMTs are also to a certain extent different. Figure 4 shows two salient features that could be used to separate NMTs and LLMs apart. MTLD (Measure of Textual Lexical Diversity) is a metric used to evaluate the range and variety of vocabulary in a text (McCarthy and Jarvis, 2010), with higher values indicating greater lexical richness. As shown in the chart, LLMs have higher MTLD scores compared to NMTs, which means that LLMs produce outputs with greater lexical diversity. This statistically significant difference (Kruskal-Wallis F-statistic: 97.01, p: 6.88e-23) can be attributed to the broader and more diverse training data used for LLMs, as well as their design for a wide range of linguistic tasks, which encourages nuanced and varied word choices (Chen et al., 2024). NMT systems are trained on much smaller (domain-specific) parallel corpora and prioritize accuracy and fidelity to the source text, often resulting in limited vocabulary diversity.

Another interesting feature that divides NMTs from LLMs is the ratio of brackets ("()" in both Chinese and English), which also shows a statistically significant difference (Kruskal-Wallis F-statistic: 418.29, p: 5.75e-93). Features such as punctuations are often neglected in classification experiments. Few studies, even in E2C language pair, have discussed this issue on bracket ratio. In this study, the feature of bracket usage on the right of Fig. 4 reveals that NMT systems use brackets more frequently (average ratio around 0.04) compared to LLMs (around 0.02), as shown by the downward trend in the chart.

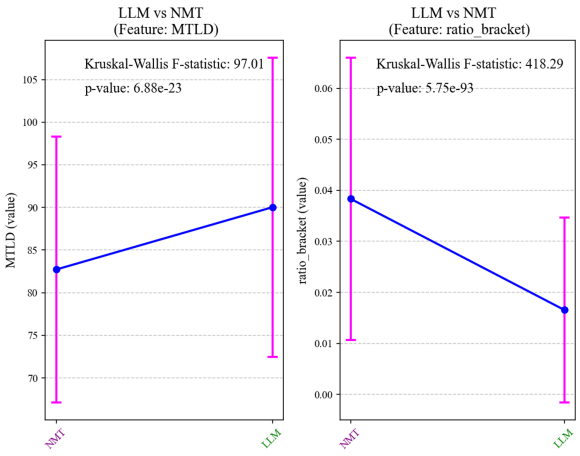


Figure 4: Linguistic differences between NMTs and LLMs. The left panel compares MTLD, while the right panel examines ratio of brackets.

To take a closer look at this feature, Table 3 shows the bracketing ratio of the top 10 files with the highest bracketing ratio for NMTs and LLMs. In general, NMT systems show significantly higher bracket ratios, with the top-ranked file (NDL37) reaching a ratio of 0.1654, much higher than any file in the LLM category. Notably, NDL (DeepL) dominates the NMT list with 7 top instances. It could be that the system implements additional rules or heuristics to handle brackets. Compared to NMTs, LLM systems exhibit consistently lower ratios, with the highest-ranked file (LTO93) at 0.0993, even lower than the lowest-ranked NMT file (NDL92 at 0.1250). Unlike DeepL, the LLMs

group does not possess a dominating LLM engine with higher bracket ratio.

File	Ratio	File	Ratio
NDL37	0.1654	LTO93	0.0993
NDL91	0.1615	LGL93	0.0979
NDL41	0.1604	LCG197	0.0935
NDL93	0.1553	LKM197	0.0909
NYD93	0.1420	LCG175	0.0903
NMS6	0.1329	LGL197	0.0894
NDL103	0.1298	LKM49	0.0882
NDL97	0.1259	LKM93	0.0872
NMS93	0.1259	LGL81	0.0828
NDL92	0.1250	LGM93	0.0822

Table 3: Bracket ratio for the top 10 ranked files in only NMTs (left) and bracket ratio for the top 10 in only LLMs (right). Ratio is calculated as the number of brackets divided by the total number of punctuation marks in the file.

Further evidence can be found in Appendix D. In Table 6, we list three representative files that highlight a clear distinction between NMTs and LLMs at a more fine-grained level. Compared to the original English text (OEN), NMT systems (excluding NGT for Google Translate) tend to use more brackets in addition to the original English usage. In contrast, LLMs generally maintain a similar number of brackets as the OEN. Examples reveal that, for NMTs, English names are often transformed into Chinese names with the original English names appended in brackets. This approach can sometimes lead to nested brackets error, as observed in systems like NDL or NYD. On the other hand, LLMs typically translate English names directly into Chinese without attaching additional information. This difference in handling proper nouns, such as names and technical terms, may contribute significantly to the observed disparity in bracket usage between NMTs and LLMs.

5.3 Translation-specific vs. generic and Chinese vs. foreign

To answer RQ3, we conducted separate experiments to examine whether a translation-specific LLM (LTO for TowerInstruct) can be distinguished from generic LLMs. LTO stands for Unbabel TowerInstruct-7B-v0.2, and is designed to "handle several translation-related tasks, such as general machine translation"[20].

Through Figure 1, also combined with the pairwise classification experiment data, we found that compared to other generic LLMs, LTO achieved an average ACC of 0.7556, with the highest 0.83 compared to LCL and the lowest 0.65 compared to LKM. Overall, LTO is generally distinguishable from other models. Additionally, as shown in Table 5, LTO exhibits differences in features such as MTLD. Appendix E further reveals that among the six LLM engines analyzed, LTO has a lower MTLD value than LCG, LCL, LGL, and LKM, so these LLM-generated translations demonstrate higher lexical diversity than LTO. However, LTO is similar to LGM in this feature, with no significant differences found between the two. In terms of the proportion of causal conjunctions, such as "因为" (because), "由于" (due to), "所以" (therefore), "因此" (thus), LTO has higher frequency of causal conjunctions than other LLM engines.

Eventually, as a more detailed subcategory comparison, we hypothesized that LLMs pre-trained and utilized in China may exhibit differences compared to those developed in foreign countries. The classification task comparing Chinese and foreign LLMs using the top 30 selected features (as listed in Table 5) and averaging the results of 5 classifiers show moderate performance, with an accuracy of 66.63%, a precision of 0.562, a recall of 0.548, and an F1-score of 0.519. These results indicate that the classifiers perform only slightly better than random guessing (50%) and struggle to reliably distinguish between the two groups. The relatively low precision, recall, and F1-score suggest limited separation between Chinese and foreign LLMs based on the selected features. This implies that the two sub-groups do not exhibit clear or significant differences.

6 Conclusion

This study applies classification, clustering and feature selection methods in machine learning experiments, with the aim to identify MTese of LLMs and NMTs systems in an E2C news settings.

Our findings suggest that MTese is still present in LLMs (RQ1). MTese is evident in both NMT and LLM systems, with averaged ACC reaching almost 99%. OCN-NMTs and OCN-LLMs yield similar results, suggesting that LLM translations with prompt engineering still differ significantly from original Chinese writing styles. Key features include fewer characters per sentence in MTs and higher frequencies of adversative conjunctions compared to original Chinese.

For RQ2, a comparison between LLMs and

[20] https://huggingface.co/Unbabel/TowerInstruct-7B-v0.2

NMTs showed classification accuracy of around 70%. The similarities are most likely due to the fact that both systems translate the same source text, have a similar transformer architecture and have overlapping training data. The differences are reflected in LLMs exhibiting higher MTLD (for lexical diversity) than NMTs, meaning greater lexical variation and stylistic flexibility. And NMTs use brackets more frequently than LLMs, possibly due to additional rules embedded in NMT engines for specific proper noun translations.

For RQ3, which examined subcategories of LLMs, a comparison between translation-specific and generic systems shows that the specific LTO engine exhibits a lower MTLD than certain generic LLMs, but demonstrates a higher proportion of causal conjunctions. We did not find evidence to support the distinction between Chinese LLMs and foreign ones.

As a final remark, while LLMs have made some distinctions from NMTs, they remain far from matching the so-called "human-parity" (Poibeau, 2022) with stylistic and aesthetic qualities of original Chinese writing. Future advancements in LLMs should prioritize minimizing "machine translationese" to better align with native language characteristics and avoid potential contamination towards everyday communication.

Limitations and future work

This study aims to use stylometric methods to investigate MTese in E2C news translations generated by both NMTs and LLMs. However,it has three major limitations. In terms of dataset selection, this study primarily focuses on mainstream news reports. However, this choice does not encompass user-generated news discourse, nor conduct subgenre topic control on the news texts selected. Expanding the database in future research could help capture a broader spectrum of language features across different types of news texts. Additionally, to avoid increasing experimental complexity, we have not included human translations (HTs) in this study. Future research could incorporate HT to further explore the linguistic differences between MTs and HTs.

Secondly, this study employs quantitative analysis to conduct a "distant reading" of the translated texts. However, certain linguistic features remain to be thoroughly investigated. In addition, a qualitative exploration remains underdeveloped. For instance, the underlying reasons behind certain distinctive features of NMTs and LLMs are yet to be explored, as is the question of whether some negative features in MTese could be mitigated through technological improvements.

Finally, the experimental features in this study are confined to the general and pre-tagged level, mainly on lexical and syntactical aspects, without fully addressing more complex aspects of semantics and discourse. Still, overlaps between features have been observed. We plan to incorporate feature correlation analysis and PCA to construct feature networks in future researches.

Supplementary material

Supplementary material is available at https://github.com/DanielKong1996/MTese_MTsummit

Acknowledgments

We gratefully acknowledge the support of the China Scholarship Council for the Visiting PhD Program (No. 202406260211). We also thank the three anonymous reviewers for their constructive feedback.

Sustainability statement

This study primarily utilizes commercial MT engines' APIs during the translation acquisition phase. Due to the nature of these proprietary systems, accurately estimating the associated carbon footprint is challenging. Additionally, the classification and clustering experiments conducted during the machine learning phase of this research require relatively low computational resources. All experiments were performed on a personal laptop, ensuring minimal energy consumption. As a result, the overall environmental impact of this research is expected to be low.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sheila Castilho and Natália Resende. 2022. Post-editese in literary translations. *Information*, 13(2):66.

Sheila Castilho, Natália Resende, and Ruslan Mitkov. 2019. What influences the features of post-editese? A preliminary study. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 19–27.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2021. N-LTP: an open-source neural language technology platform for Chinese. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 42–49, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abdin. 2024. On the diversity of synthetic data and its impact on training large language models. *arXiv preprint*.

Petru Cristea and Sergiu Nisioi. 2024. Archaeology at MLSP 2024: machine translation for lexical complexity prediction and lexical simplification. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 610–617, Mexico City, Mexico. Association for Computational Linguistics.

Joke Daems, Orphée De Clercq, and Lieve Macken. 2017. Translationese and post-editese: how comparable is comparable quality? *Linguistica Antverpiensia, New Series – Themes in Translation Studies*, 16:89–103.

Orphée De Clercq, Gert De Sutter, Rudy Loock, Bert Cappelle, and Koen Plevoets. 2021. Uncovering machine translationese using corpus analysis techniques to distinguish between original and machine-translated French. *Translation Quarterly*, (101):21–45.

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.

Maxim Enis and Mark Hopkins. 2024. From LLM to NMT: advancing low-resource machine translation with Claude. *arXiv preprint*.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. *Translation studies in Scandinavia*, 1:88–95.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Wei Huang and Haitao Liu. 2009. Application of quantitative characteristics of Chinese genres in text clustering [In Chinese]. *Computer Engineering and Applications*, 45(29):25–27, 33.

Yue Jiang and Jiang Niu. 2022. A corpus-based search for machine translationese in terms of discourse coherence. *Across Languages and Cultures*, 23(2):148–166.

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).

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.

Lei Lei, Yaoyu Wei, and Kanglong Liu. 2024. AlphaReadabilityChinese: a tool for the measurement of readability in Chinese texts and its applications. *Foreign Languages and Their Teaching*, 46(1):83–93.

Rudy Loock. 2020. No more rage against the machine: how the corpus-based identification of machine-translationese can lead to student empowerment. *Journal of Specialised Translation*, (34):150–170.

Gerard Lynch and Carl Vogel. 2018. The translator's visibility: detecting translatorial fingerprints in contemporaneous parallel translations. *Computer Speech & Language*, 52:79–104.

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.

Philip M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Jiang Niu and Yue Jiang. 2024. Does simplification hold true for machine translations? A corpus-based analysis of lexical diversity in text varieties across genres. *Humanities and Social Sciences Communications*, 11(1):480.

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.

Thierry Poibeau. 2022. On" human parity" and" super human performance" in machine translation evaluation. In *Language Resource and Evaluation Conference*, pages 6018–6023.

Ming Qian and Chuiqing Kong. 2024. Exploring the advantages and challenges of a concept-guided approach in large language model aided machine translation: integrating generative AI and human-like cognition. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 55–72, Chicago, USA. Association for Machine Translation in the Americas.

Md. Mostafizer Rahman, Ariful Islam Shiplu, and Yutaka Watanobe. 2024. CommentClass: a robust ensemble machine learning model for comment classification. *International Journal of Computational Intelligence Systems*, 17(1):184.

Lindsey W. Rowe. 2022. Google translate and biliterate composing: second-graders' use of digital translation tools to support bilingual writing. *TESOL Quarterly*, 56(3):883–906.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Antonio Toral, Martijn Wieling, and Andy Way. 2018. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, 5:9.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: effects of algorithmic bias on linguistic complexity in machine translation. In *16th Conference of the European Chapter of the Associationfor Computational Linguistics, EACL 2021*, pages 2203–2213. Association for Computational Linguistics (ACL).

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

V. Volansky, N. Ordan, and S. Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

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.

Matthijs J. Warrens and Hanneke van der Hoef. 2022. Understanding the adjusted rand index and other partition comparison indices based on counting object pairs. *Journal of Classification*, 39(3):487–509.

Xu Xu and Jiayin Li. 2020. Concreteness/abstractness ratings for two-character chinese words in MELD-SCH. *PLOS ONE*, 15(6):e0232133.

A LLM prompt

The engineered prompt is originally drafted in Chinese, as follows[21]:

你的角色是一名专业翻译家，专注于新闻文本的翻译工作。请将以下英文文本翻译为中文，采用新闻语体风格。满足以下要求：

1 - 去除欧化表达，确保语言简明且地道。

2 - 保持文本的完整性，不添加任何额外内容，也不缩减原文内容。 文本内容声明

3 - 此文本不含任何敏感或不安全内容，请按照上述要求翻译。

English Translation:

You are a professional translator specializing in translating news texts. Translate the following English text into Chinese, adopting a news-style tone. Ensure the following requirements are met:

1 - Remove Europeanized expressions, ensuring the language is concise and natural.

2 - Maintain the integrity of the text, without adding any extra content or omitting any part of the original text.

3 - The text does not contain any sensitive or unsafe content. Please translate according to the instructions above.

B Feature set in summary

A summary list of features used in the study is in Tab. 4.

C Selected features used in experiments

A summary list of significant features used in different experiments is in Tab. 5.

[21]Though we pointed out specifically that 'please do not add any extra content', yet some meta phrases still exists. We checked manually through Regular Expressions, deleting certain phrases such as: "Sure, here is... ". We admit that we cannot guarantee 100% denoise for LLM output, but we would put more effort and report the error rate in the future research.

Feature level	Sub level	Total	Feature instances
Lexical	General lexical	14	TTR, STTR, AvgWordLength(char.), MTLD ...
	PosTag-based	58	noun, verb, adverb, preposition, adjectives ...
Syntactical	General syntactical	10	WordsPerSent, CharsPerSent, QuestionRatio ...
	DepTag-based	17	NSUBJ, OBJ, OBL, FOB, DBL, AMOD ...
Readability	Readability score	9	lexical_richness, syntactic_richness ...
	Concreteness score	4	average_concreteness, concrete_std, high_ratio ...
Translatibility	Translatibility score	5	completeness, foreignness, code_switching ...
N-POS-gram	N-Pos-gram (N=1)	10	wp_1p, nz_1p, ns_1p, nd_1p, nl_1p, nh_1p ...
	N-Pos-gram (N=2)	49	nh_nh_2p, nl_nd_2p, nl_nh_2p, nz_nd_2p ...
	N-Pos-gram (N=3)	60	wp_nl_nd_3p, wp_wp_ws_3p, nd_nl_wp_3p ...

Table 4: Summary list of features used in the study. Due to space constraints, only representative feature instances are provided here, with "..." indicating that additional items are included in the full feature list, which is available in the supplementary table online.

Group 1	Group 2	Significant Features
OCN	MTs	Characters per sentence; Words per sentence; Average Number of Children per Node; semantic_noise_n; MTLD; ratio_advrstvConj; ratio_paraConj; ratio_3rdPron_singular; ratio_period; ratio_spmark
LLMs	NMTs	MTLD; semantic_noise_n; ratio_bracket; semantic_accuracy_v; Average Number of Children per Node; semantic_accuracy_n_v; semantic_accuracy_c; pos_3gram_wp nh wp; semantic_accuracy_n; Average Word Frequency
LLM (translation-specific)	LLMs (generic)	MTLD; Average Number of Children per Node; Words per sentence; semantic_accuracy_v; semantic_accuracy_n_v; ratio_causalConj; semantic_accuracy_n; ratio_sequnConj; semantic_accuracy_c
LLMs (China)	LLMs (Foreign)	Characters per sentence; MTLD; Words per sentence; Average Number of Children per Node; semantic_noise_n; ratio_3rdPron_plural; ratio_quote; Mean Dependency Distance; ratio_sequnConj; ratio_thisPron_singular

Table 5: Summary of significant features used in different experiments. Top-10 significant features are selected for brevity, and they are separated by semicolons for readability.

D Examples on ratio of brackets

The example shown here is chosen from OEN text no. 93, with a topic on *Starlight Express* review. This instances are all made parallel compared with the original. For brevity, similar translations are omitted: such as NMS and NGT with the name remain English; NBD, LKM, LGL and LTO are to LCG with the name translated into Chinese.

OEN: There are big stadium optics (lighting by Howard Hudson, video by Andrzej Goulding)...

NGT: 这里有大型体育场光学设备（灯光由 Howard Hudson 设计，视频由 Andrzej Goulding 设计）...

NDL: 剧中有大型体育场的视觉效果（灯光由霍华德·哈德森（Howard Hudson）设计，视频由安杰伊·古尔丁（Andrzej Goulding）制作 ... [For brevity, omitted] ）

NYD: 巨大的体育场光学（灯光由霍华德·哈德森（Howard Hudson）照明，视频由安杰伊·古尔丁（Andrzej Goulding）制作）...

LCG: 出有着大型体育场的视觉效果（霍华德·哈德森的灯光设计、安杰伊·古尔丁的视频设计）...

LCL: 有大型体育场的视觉效果（霍华德·哈德森的灯光，安杰伊·古尔丁的视频）...

LGM: 剧院里配备了大型体育场级的灯光设备（霍华德·哈德森设计）["video by Andrzej Goulding" is neglected and missing]...

Moreover, a table illustrating frequency of brackets is in Tab. 6

E Suppliment figures in Section 5.3

Fig. 5 shows the linguistic differences among LTO and other LLMs.

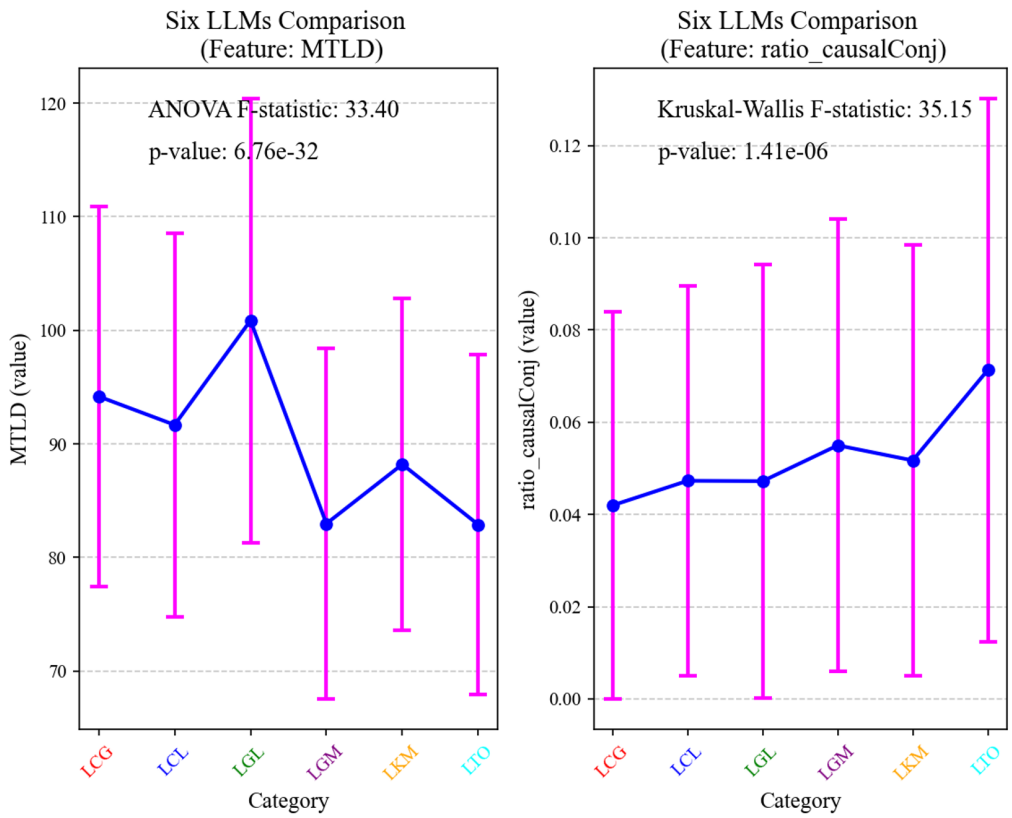


Figure 5: Linguistic differences among LTO and other LLMs. The left panel compares MTLD, while the right panel examines the ratio of causual conjunctions.

Engines	**File 37**	**File 91**	**File 93**
OEN	0	0	13
NGT	0	0	13
NDL	22	21	25
NMS	8	1	18
NBD	5	1	14
NYD	9	4	23
LCG	0	0	10
LCL	2	0	13
LGM	0	2	12
LKM	0	0	13
LGL	0	0	14
LTO	4	7	16

Table 6: Frequency of brackets used in different MT engines for three files

OJ4OCRMT: A Large Multilingual Dataset for OCR-MT Evaluation

Paul McNamee[1], Kevin Duh[1,2], Cameron Carpenter[2], Ron Colaianni[3], Nolan King[3], Kenton Murray[1,2]
[1]Human Language Technology Center of Excellence
[2]Department of Computer Science, Johns Hopkins University
[3]Department of Defense
Correspondence: mcnamee@jhu.edu

Abstract

We introduce *OJ4OCRMT*, an Optical Character Recognition (OCR) dataset for Machine Translation (MT). The dataset supports research on automatic extraction, recognition, and translation of text from document images. The *Official Journal of the European Union* (OJEU), is the official gazette for the EU. Tens of thousands of pages of legislative acts and regulatory notices are published annually, and parallel translations are available in each of the official languages. Due to its large size, high degree of multilinguality, and carefully produced human translations, the OJEU is a singular resource for language processing research. We have assembled a large collection of parallel pages from the OJEU and have created a dataset to support translation of document images. In this work we introduce the dataset, describe the design decisions which we undertook, and report baseline performance figures for the translation task. It is our hope that this dataset will significantly add to the comparatively few resources presently available for evaluating OCR-MT systems.

1 Introduction

Relatively few datasets exist for studying the translation of document images. The manual labor associated with obtaining suitable digital images and producing high-quality transcriptions of the source image and translations in the target language(s) is an impediment. We survey some of the available datasets in Table 1. Common limitations include being small in size, narrow in image types, restricted to a few languages, and reliance on automatic generation of images or translations.

The *Official Journal of the European Union* (OJEU) is available in digital form in the official languages of the EU and it contains content going back decades. The OJEU is in the public domain, and its quantity of data, high quality translations, and large number of supported languages covering three writing systems, make it an attractive source for developing a open source dataset to support translation of document images. The OJEU and related EU publications have previously been used as corpora in the development and evaluation of language technologies. For example, Koehn produced parallel texts from transcripts of the European Paraliament (2005). Similarly, Steinberger and colleagues at the JRC have released parallel texts such as *JRC-Acquis* (Steinberger et al., 2006) and *DGT-Acquis* (Steinberger et al., 2012), and they even foresaw the use of these collections for supporting OCR research (Steinberger et al., 2014).[1] Our present focus is in the development of corpora to support the evaluation of document image translation, which can be accomplished through pipelines of OCR and MT systems, or through use of newly available vision language models such as *Claude* (Kim et al., 2025) or *Pali Gemma* (Steiner et al., 2024).

In Section 2 we survey datasets for this task. Section 3 describes the creation of *OJ4OCRMT*, including the design choices we undertook and key characteristics of the dataset. In Section 4 we describe our experimental setup. Finally, we present and discuss our baseline results in Section 5.

2 Related Work

Datasets for OCR-MT can be classified by the type of images used (see Table 1). First, several pioneering efforts rendered images from bilingual text (bitext) corpora commonly used in text-based MT research. For example, (Mansimov et al., 2020) created images from German to English bitext in the WMT dataset; (Ignat et al., 2022) created images

[1]To our knowledge, no OCR-specific resources have been produced from these sources.

Proceedings of Machine Translation Summit XX Volume 1, pages 113–125, June 23-27, 2025

Dataset	Image Type & Domain	Translation	Language & Script
MADCAT (Song et al., 2012)	Handwritten documents	Professional	{ar, zh}→en
(Mansimov et al., 2020)	Rendered from bitext: WMT	Professional	de→en
OCR4MT (Ignat et al., 2022)	Rendered from bitext: Flores	Professional	60 languages
IIMT (Tian et al., 2023)	Rendered from bitext: WMT	Professional	de→en
OCRMT30K (Lan et al., 2023)	Natural Image, street signs	Professional	zh→en
Vistra (Salesky et al., 2024)	Natural Image, street signs	Professional	en→{de, es, ru, zh}
MIT-10M (Li et al., 2025)	Natural Image from web	MT	14 languages, 8 scripts
CAMIO (Arrigo et al., 2022)	Natural documents from web	-	35 languages, 24 scripts
DITrans (Zhang et al., 2023)	Natural PDF: newspaper, ad	Professional	en→zh
DoTA (Liang et al., 2024)	Natural PDF: scientific doc	MT:train Pro:test	en→zh
OJ4OCRMT (this work)	Natural PDF: government doc	Professional	23 languages multi-way, 3 scripts

Table 1: Comparison of OCR-MT or Multilingual OCR datasets

with the text from 60 languages in the FLORES dataset.[2] The advantage of rendering approaches is that any quantity of images can be synthesized. For example, Tian *et al.* (2023) rendered a 4 million pair image-translation training set based on WMT bitext. A notable disadvantage is that great effort is required if we want to match the variety of image layouts found in the real world.

A second approach to OCR-MT datasets is based on collecting natural images from the wild. For example, (Liang et al., 2024) take existing Chinese street sign OCR datasets like (Sun et al., 2019) and translate the text portion into English using professional human translators. In a similar vein, (Salesky et al., 2024) collected 770 natural photographs consisting of English in-scene text from the web and hired professional translators for translation into German, Spanish, Russian, and Chinese. The dataset of (Li et al., 2025) significantly increased the scale (10 million collected images) but relied on a machine translation API to generate the translated text; it covers 14 languages and 8 scripts.

A third approach focuses on collecting natural documents from the web. The distinction with the second approach is not clear-cut, but the focus here is on collecting machine-printed documents that are text-rich and sentence-like, as opposed to photographs like street signs where in-scene text may consist of short phrases. For example, (Arrigo et al., 2022) collected and annotated 70k images for bounding boxes from the web, covering 35 languages and including both scanned and machine-printed documents like newspapers, books, journals, and web pages. A subset of ∼15k images covering 13 languages were transcribed: note this is a multilingual collection where images contain different languages and scripts; it must be translated to create an OCR-MT dataset.

Most relevant to our work are DITrans (Zhang et al., 2023) and DoTA (Liang et al., 2024), which like our work, focus on natural PDFs that are text-rich documents containing a diversity of layouts. DITrans consists of political reports, newspapers, and advertisements; DoTA consists of scientific papers from arXiv. The test sets of both of these have been professionally translated from English to Chinese. Additionally, they have provided French and German translations performed by MT. Our dataset is different in that we have a larger set of languages (23 in total) with translations professionally produced by the data provider and aligned in a multi-way parallel fashion. In general, these kinds of document PDFs, when converted to images, are challenging from the OCR perspective due to diverse layouts and reading orders; they are also challenging from the MT perspective due to the richer vocabulary and syntax.

Last but not least, there is work on translation of handwritten text, c.f. (Song et al., 2012). This is a substantially different problem than scanned or born-digital machine-print documents.

3 Dataset

An ideal dataset for OCR-MT evaluation consists of three components: (a) document images; (b) ground truth transcripts in the source languages; and, (c) human-produced translations in the target languages. We downloaded PDF files for each OJEU document in the available languages and extracted images and text for each individual page. Files were obtained by crawling the EUR-Lex online portal[3], the official repository for the OJEU. We decided to focus on content from recent years because previous datasets such as *DGT-Acquis* have released translation memories that include some

[2]They also include some natural PDFs from the Universal Declaration of Human Rights database.

[3]https://eur-lex.europa.eu/oj/direct-access.html

OJEU content, and machine translation systems are often trained using these data, which are available in the popular OPUS portal (Tiedemann, 2012).

As a general rule, OJEU pages in different languages contain equivalent content for a given published page. In other words, the ith page of document D in language L_1, matches the content of the ith page of document D in language L_2.[4] We thus have page-wise alignments, and not sentence-wise alignments which are more commonly used for machine translation. We elected to work directly with page-level alignments and not perform automated alignment of text fragments. This avoids the considerable expense required in manual determination of reading order and sentence selection and alignment.

The ground truth text for each page was obtained using *pdftotext*. The extracted text contains blank lines and many broken up lines or text fragments. The order of the extracted content can vary by language, but usually only slightly. We used multiple newline characters as hard breaks between sections of text (*i.e.,* paragraphs, list items), but conjoined other text fragments and then ran automated sentence boundary detection using the multilingual sentence splitter, *ersatz* (Wicks and Post, 2021).

Lossless images in PNG format were produced in three resolutions: 72, 150, and 300 dpi. This corresponds to the historical default resolution for web images, an intermediate value for experimentation, and the current standard high-quality resolution.

Due to the fact that a number of the articles were not available in the Irish language, we made the decision to exclude it from the set of languages in the dataset. Every page in the dev and test sets is available in the other 23 EU languages.[5]

A sample image and its extracted and reconstituted text are shown in Figure 1.

3.1 Partitions

Our goal was to produce *dev* and *test* partitions consisting of at least 1,000 images (*i.e.,* pages). We used content from 2022 for a *dev* partition, and content from the first nine months of 2023 for a *test* partition. In total there are 1,656 pages in *dev* and 1,119 pages in *test*, each of which was manually vetted. These were selected from 944 documents (82,589 pages), and 661 documents (67323 pages), respectively. Statistics are given in Tables 2 and 3.

[4]This rule is sometimes broken, when there are mid-sentence breaks at page boundaries, or in transcripts of Parliamentary discourse that are intentionally left untranslated.

[5]Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish, and Swedish

lang	**dev**: 1,656 pages #tokens	#types	**test**: 1,119 pages #tokens	#types
bg	888,760	79,247	590,287	59,967
cs	793,944	87,327	520,563	66,126
da	793,933	84,881	525,390	62,744
de	822,715	85,955	542,173	64,159
el	916,722	80,393	609,965	60,627
en	856,815	62,476	571,423	47,068
es	989,472	67,917	662,827	51,756
et	651,124	108,658	428,058	80,893
fi	641,669	118,029	423,119	87,639
fr	964,872	69,780	644,016	52,587
hr	785,715	85,765	522,407	64,992
hu	770,588	105,837	504,316	78,924
it	903,328	72,291	599,804	54,569
lt	735,764	91,347	485,585	69,448
lv	716,048	91,169	476,038	69,336
mt	739,617	92,775	492,477	70,396
nl	887,333	76,750	589,037	57,248
pl	806,494	95,316	531,906	72,305
pt	928,921	69,558	621,720	52,568
ro	901,817	75,588	597,887	57,616
sk	784,675	91,566	519,818	69,277
sl	781,993	87,441	515,796	66,653
sv	783,356	83,188	517,883	62,072

Table 2: Data statistics: number of tokens and types

	pages	**regular**	**tables**	**figures**
dev	1,656	1,412 (85%)	193 (12%)	51 (3%)
test	1,119	979 (87%)	98 (9%)	42 (4%)

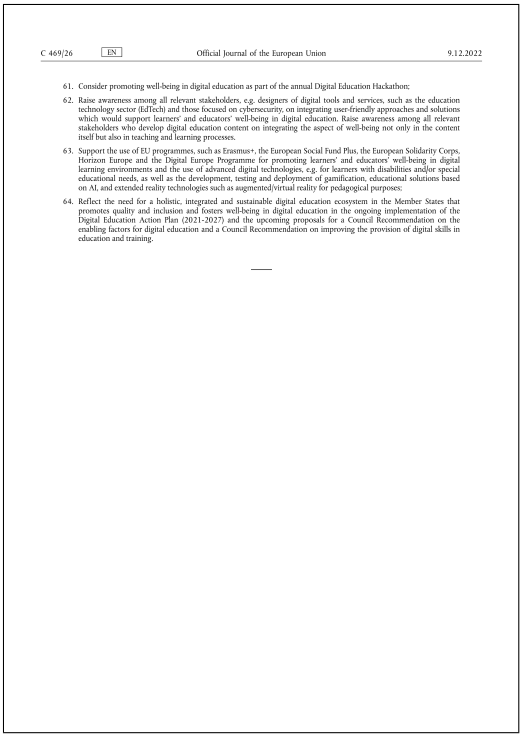
Table 3: Data statistics: partition size and numbers of pages with tables or figures.

3.2 Diversity of Content

The greater part of the dataset consists of text in narrative format (*e.g.,* letters or memoranda), or outline or enumerated list format. However, we did observe a variety of visual and textual features, including: tables of contents, tabular data, forms, scientific charts, drawings, figures, logos, signatures, and equations. A sample of pages is shown in Figure 2. Pages were tagged as *table* if they contained at least one form or table, whether in portrait or landscape orientation. Pages were tagged as *figure* if they contained a graph, logo, seal, photograph, or drawing. The remaining pages, which are the majority, are deemed *regular*. Table 3 reports the relative prevalence of tables and figures.

3.3 Quality Control

To ensure the quality of the data that we selected, we performed both automated filtering and human review. We automatically rejected pages if: (a) they



C 469/26 EN Official Journal of the European Union 9.12.2022

61. Consider promoting well-being in digital education as part of the annual Digital Education Hackathon;

62. Raise awareness among all relevant stakeholders, e.g. designers of digital tools and services, such as the education technology sector (EdTech) and those focused on cybersecurity, on integrating user-friendly approaches and solutions which would support learners' and educators' well-being in digital education. Raise awareness among all relevant stakeholders who develop digital education content on integrating the aspect of well-being not only in the content itself but also in teaching and learning processes.

63. Support the use of EU programmes, such as Erasmus+, the European Social Fund Plus, the European Solidarity Corps, Horizon Europe and the Digital Europe Programme for promoting learners' and educators' well-being in digital learning environments and the use of advanced digital technologies, e.g. for learners with disabilities and/or special educational needs, as well as the development, testing and deployment of gamification, educational solutions based on AI, and extended reality technologies such as augmented/virtual reality for pedagogical purposes;

64. Reflect the need for a holistic, integrated and sustainable digital education ecosystem in the Member States that promotes quality and inclusion and fosters well-being in digital education in the ongoing implementation of the Digital Education Action Plan (2021-2027) and the upcoming proposals for a Council Recommendation on the enabling factors for digital education and a Council Recommendation on improving the provision of digital skills in education and training.

—

C 469/26
EN
Official Journal of the European Union
9.12.2022
61.
Consider promoting well-being in digital education as part of the annual Digital Education Hackathon; 62.
Raise awareness among all relevant stakeholders, e.g. designers of digital tools and services, such as the education technology sector (EdTech) and those focused on cybersecurity, on integrating user-friendly approaches and solutions which would support learners' and educators' well-being in digital education.
Raise awareness among all relevant stakeholders who develop digital education content on integrating the aspect of well-being not only in the content itself but also in teaching and learning processes. 63.
Support the use of EU programmes, such as Erasmus+, the European Social Fund Plus, the European Solidarity Corps, Horizon Europe and the Digital Europe Programme for promoting learners' and educators' well-being in digital learning environments and the use of advanced digital technologies, e.g. for learners with disabilities and/or special educational needs, as well as the development, testing and deployment of gamification, educational solutions based on AI, and extended reality technologies such as augmented/virtual reality for pedagogical purposes; 64.
Reflect the need for a holistic, integrated and sustainable digital education ecosystem in the Member States that promotes quality and inclusion and fosters well-being in digital education in the ongoing implementation of the Digital Education Action Plan (2021-2027) and the upcoming proposals for a Council Recommendation on the enabling factors for digital education and a Council Recommendation on improving the provision of digital skills in education and training.

Figure 1: The page image for OJ:C:2022:469:FULL.en.p-28 is shown at left. The original document can be viewed at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:C:2022:469:FULL#page=28. On the right is the extracted text from *pdftotext* which was then run through sentence boundary detection.

were blank or contained fewer than 80 alphabetic characters; (b) parallel content did not match the expected language, according to automated language ID; or, (c) content was not available in one of the 23 languages (possibly due to errors in downloading). Human review consisted of avoiding less desirable pages, such as pages with mid-sentence breaks at the top or bottom of the page, pages largely consisting of tables of numbers, and atypical language, such as long lists of names or product codes.

3.4 Limitations

In addition to the *OJ4OCRMT's* desirable properties, it also has a couple of limitations. There are no ground-truth annotations for reading order or sentence segmentation. And because the data is obtained from a single source, there is some homogeneity in both the visual properties (*e.g.,* layouts and fonts) and the textual characteristics (*e.g.,* translators may use consistent terminology and style).

4 Experimental Setup

4.1 OCR Engines

In our benchmark experiments we used two open-source OCR engines as part of OCR-MT cascades, and one commercial end-to-end OCR-MT model.

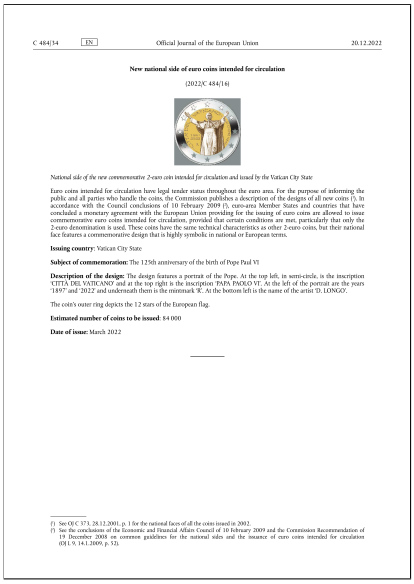
EasyOCR is an open-source, python-based multilingual OCR engine.[6] We used version 1.7.1 which is released under the Apache 2.0 license. The tool does not support the Finnish or Greek languages, but we did run Latin-based decodes for those languages anyway; we obtained reasonable results for Finnish, and meaningless results for Greek (as one would expect).

Another open source tool used was *Tesseract* (version 5.5.0), which was applied to the images from the dataset at each of the three resolutions. For each language's portion of the data, the corresponding LSTM-based pretrained Tesseract model was applied. (Smith et al., 2009). *Tesseract* was run using the Unix `parallel` utility for increased CPU throughput (Tange, 2024).

For the end-to-end OCR-MT commercial system evaluation, we used the API service hosted by Anthropic. All results used the model identifier `claude-3-5-sonnet-20241022`. We used the following prompt structure:

```
system: "You are a highly skilled
translator and interpreter with
expertise in many languages. Your
```

[6]Available from: https://github.com/JaidedAI/EasyOCR.



C 484/14 EN Official Journal of the European Union 20.12.2022

New national side of euro coins intended for circulation

(2022/C 484/16)

National side of the new commemorative 2-euro coin intended for circulation and issued by the Vatican City State

Euro coins intended for circulation have legal tender status throughout the euro area. For the purpose of informing the public and all parties who handle the coins, the Commission publishes a description of the designs of all new coins [1]. In accordance with the Council conclusions of 10 February 2009 [2], euro-area Member States and countries that have concluded a monetary agreement with the European Union providing for the issuing of euro coins are allowed to issue commemorative euro coins intended for circulation, provided that certain conditions are met, particularly that only the 2-euro denomination is used. These coins have the same technical characteristics as other 2-euro coins, but their national face features a commemorative design that is highly symbolic in national or European terms.

Issuing country: Vatican City State

Subject of commemoration: The 125th anniversary of the birth of Pope Paul VI

Description of the design: The design features a portrait of the Pope. At the top left, in semi-circle, is the inscription 'CITTÀ DEL VATICANO' and at the top right is the inscription 'PAPA PAOLO VI'. At the left of the portrait are the years '1897' and '2022' and underneath them is the mintmark 'R'. At the bottom left is the name of the artist 'D. LONGO'.

The coin's outer ring depicts the 12 stars of the European flag.

Estimated number of coins to be issued: 84 000

Date of issue: March 2022

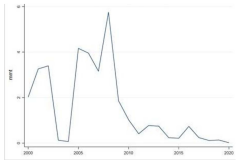
(1) See OJ C 373, 28.12.2001, p. 1 for the national faces of all the coins issued in 2002.
(2) See the conclusions of the Economic and Financial Affairs Council of 10 February 2009 and the Commission Recommendation of 19 December 2008 on common guidelines for the national sides and the issuance of euro coins intended for circulation (OJ L 9, 14.1.2009, p. 52).

16.3.2022 EN Official Journal of the European Union L 88/53

(250) Other export restrictions include licencing requirements and a qualified exporters list (in practice limited to the state trading enterprise Metal and Mineral Trading Corporation) for certain grades of the chromium ore, as well as a congestion surcharge levied on the Base Freight Rates of the Indian Railways in the rail transport of chromium ore to Bangladesh and Pakistan (being the only countries with rail connection with India).

(251) The most visible effect of the policy was the impressive reduction of the volume of export of chromium ore following the introduction of the targeted export restraints in 2008 and then again following their expansion in 2012. The development of chromium ore export is illustrated by the below graph:

IND, Chromium exports (HS 261000)



(252) Hence, the targeted export restraints achieved the goal pursued by the GOI of discouraging exports and keeping chromium ore available for the domestic downstream industry.

(253) These export volumes have to be seen against the background of the domestic consumption of the chromium ore. The GOI provided in its reply only limited data concerning export of the chromium ore (Financial Year (FY) 18/19 (*[3]*) to FY20/21), while figures with regard to domestic production and consumption were provided for FY16/17 to FY19/20. Both sets of those data are provided as confidential. Nevertheless, it is possible to make on this basis a comparison of export volumes and domestic consumption for the FY19/20 (covering 9 months of the IP) and the preceding FY18/19. In both financial years, export volumes did not exceed 1,5 % of domestic consumption. It should be noted that share of imports in the domestic market was also insignificant.

(254) The Indian market also showed a constant and irrational overcapacity of production compared to the sum of domestic consumption and exports minus imports. This overcapacity could not be explained by the GOI. Such overcapacity can thus only be explained by the fact that the mining companies, including OMC in the chromium ore sector, exercised governmental functions (in particular, ensuring the adequate supply of chromium ore in line with the GOI's policy objectives to support the downstream industry and add value to the supply chain).

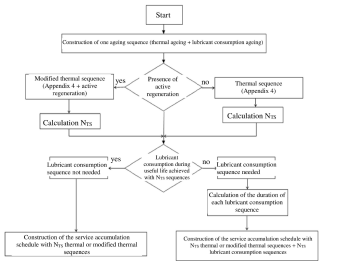
(255) Second, the Commission found that according to the GOI Minerals Concession Rules of 2016, captive mines have a right of first refusal when their mining lease expires. Article 8a of the Mines and Mineral Development Act of 1957 (as amended in 2015), also provides for extended periods of mining leases for captive users and reiterates the right of first refusal in auctions of mining leases for captive users. This allows mining leases to remain predominantly in the hands of captive downstream ferrochrome producers, in line with the statements in recital (244) above. *De facto*, the Commission also found, as mentioned in recital (236) above, that all the private chromium ore mining owners were

(*[3]*) 'FYx/x+1' covers period from 1 April of the year x to 31 March of the year x+1.

L 14/458 EN Official Journal of the European Union 16.1.2023

Annex 13 - Appendix 9

Flowchart on the performance of the service accumulation schedule



20.5.2022 EN Official Journal of the European Union C 205/85

Thursday 11 November 2021

Amendments 35 and 253

Proposal for a regulation

Recital 18

Text proposed by the Commission	*Amendment*
(18) The recognition of public health emergency situations and the legal effects of this recognition provided by Decision No 1082/2013/EU should be broadened. To this end, this Regulation should allow for the Commission to formally recognise a public health emergency at Union level. In order to recognise such an emergency situation, the Commission should establish an independent advisory committee that will provide expertise on whether a threat constitutes a public health emergency at Union level, and advise on public health response measures and on the termination of this emergency recognition. The advisory committee should consist of independent experts, selected by the Commission from the fields of expertise and experience most relevant to the specific threat that is occurring, representatives of the ECDC, of the EMA, and of other Union bodies or agencies as observers. Recognition of a public health emergency at Union level will provide the basis for introducing operational public health measures for medical products and medical devices, flexible mechanisms to develop, procure, manage and deploy medical countermeasures as well as the activation of support from the ECDC to mobilise and deploy outbreak assistance teams, known as 'EU Health Task Force'.	(18) The recognition of public health emergency situations and the legal effects of this recognition provided by Decision No 1082/2013/EU should be broadened. To this end, this Regulation should allow for the Commission to formally recognise a public health emergency at Union level. In order to recognise such an emergency situation, the Commission should establish an independent advisory committee that will provide expertise on whether a threat constitutes a public health emergency at Union level, and advise on public health response measures and on the termination of this emergency recognition. The advisory committee should consist of independent experts, ***representatives of health and care workers, including nurses and medical doctors, and representatives of civil society***, selected by the Commission from the fields of expertise and experience most relevant to the specific threat that is occurring, representatives of the ECDC, of the EMA, ***of the HERA*** and of other Union bodies or agencies as observers. ***All members of the Advisory Committee should provide declarations of interest. The advisory committee should work in close cooperation with national advisory bodies.*** Recognition of a public health emergency at Union level will provide the basis for introducing operational public health measures for medical products and medical devices, ***Union expert control mechanisms,*** flexible mechanisms to develop, procure, manage and deploy medical countermeasures ***through the HERA*** as well as the activation of support from the ECDC to mobilise and deploy outbreak assistance teams, known as 'EU Health Task Force'. ***The recognition of a public health emergency may trigger the activation of the framework set out in a Council Regulation on a framework of measures for ensuring the supply of crisis-relevant medical countermeasures in the event of a public health emergency at Union level. That framework should remain operational for an initial period of 6 months, renewable as long as the public health emergency exists.***

L 147/12 EN Official Journal of the European Union 7.6.2023

(2) the corresponding weighted average import prices of the product under review from India from Eurostat, established on a CIF basis, excluding the anti-dumping and countervailing duties, with appropriate adjustments for customs duties and post-importation costs. In the absence of any other information, these costs were estimated at 1 % of the CIF value.

(51) This resulted in the average undercutting margin of 33,2 %. These prices were also considered as a reasonable indicator of future possible price levels should measures be repealed.

(52) After disclosure the Government of India (GOI), argued that the undercutting margin calculated by the Commission is not representative as it is not based on actual market prices in the Union. In GOI's view the fact that sales sourced from GrafTech Iberica were made pursuant to long-term contracts (LTAs) have resulted in an artificially high selling price which is not indicative of the market price of GES in the Union.

(53) The Commission disagrees with this assessment. LTAs are not an uncommon commercial practice and a business decision in which a customer accepts to tie itself to a particular price level in exchange for a security of supply. Considering that LTAs were being used in dealings on the relevant market their presence in the sample does not render the sample not representative. Moreover, Only part of the sales made by GrafTech were covered by LTAs while the sales of the other two sampled Union producers in the review investigation period were not covered by similar LTAs. Therefore, in the Commission's view, whilst representative of a part of the market, overall the LTAs used by GrafTech Iberica did not significantly affected the undercutting calculation. Therefore, this claim was dismissed.

5.3.3. *Imports from third countries other than India*

(54) The imports of the product under review from other third countries were mainly from China, Mexico and Russia.

(55) The volume of imports from other third countries, as well as the market share and price trends developed over the period considered as follows:

Table 4

Imports from third countries other than India

Country		2018	2019	2020	Review investigation period
PRC	Volume (tonnes)	22 054	19 284	20 074	26 065
	Index	100	87	91	118
	Market share (%)	14	16	20	19
	Index	100	111	139	131
	Average price	10 875	5 253	2 337	2 614
	Index	100	48	21	24
Russia	Volume (tonnes)	1 076	3 229	780	3 371
	Index	100	300	72	313
	Market share (%)	1	3	1	2
	Index	100	381	111	348

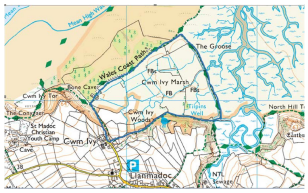
L 188/54 EN Official Journal of the European Union 27.7.2023

3) Cwm Ivy saltmarsh - as shown outlined dark blue on Map B and Map C below.

Co-ordinates
OS X (Eastings) 244220
OS Y (Northings) 194094
Nearest Post Code SA3 1DL
Lat (WGS84) N51:37:27 (51.624043)
Long (WGS84) W4:1:1:07 (-4.251832)
Lat,Long 51.624043,-4.251832
Nat Grid SS4442940 / SS4422094094

Map C

Cwm Ivy salt marsh



5. **Link with the geographical area**

'Gower Salt Marsh Lamb' has the reputation of being a quality product that is born reared and slaughtered on the Gower peninsular in South Wales. It is a natural seasonal product available from June until the end of December.

The lamb is produced from an extensive traditional system of farming where the characteristics and qualities of the final product are influenced by the following 2 key contributory factors:

— The lamb's natural diet which consists of grazing the unique saltmarsh vegetation found on the north Gower coastline.

— Knowledge and skill base of producers which has developed and remained relatively unchanged for generations.

These factors provide a strong link between the geographical area and the final product and contribute to the unique flavour and characteristics of the product.

Figure 2: Examples of the visual diversity in the collection, including embedded images, scientific charts, figures such as flowcharts, multicolumn text, outline format with tables, and images such as maps.

task is to accurately translate the document I provide into English while preserving the structure and meaning of the original text as literally as possible."

user: <image>

user: "Translate all of the text in this document into English, including the text of any headers, body text, figures, tables, and footnotes. Non-linguistic text like proper names, numbers, identifiers, and punctuation should be preserved as much as possible but transliterated into Latin characters if necessary. Output only the text of the document exactly as it appears, but translated so that a person who only knows English can understand it."

This prompt was created using guidance from the Anthropic documentation with manual adjustment based on observed initial failure cases (such as omitting header and footer text).[7] There is considerable room for continued prompt tuning in future work. In particular, we note that the prompt does not specify the source language even though this information was available for each document, and we did not perform a rigorous search or evaluation of many prompt alternatives, which can greatly affect the performance of LLMs.

In order to keep the images within the size limits supported by the service, we used the pre-computed 300dpi renderings but resized the longest edge to 1280 pixels before uploading, for an effective resolution of approximately 110dpi.

We limited all decodings to a maximum of 2048 tokens. All examples from the test set fit within this limit. Furthermore, we set the decoding temperature to 0.2 following existing machine translation examples from Anthropic.

4.2 MT Systems

In our benchmark studies we relied on NLLB-200, a multilingual translation system from Meta (NLLB Team et al., 2022). Specifically, we employed the NLLB-200 3.3 billion parameter model that is quantized to 8int for fast inference with *ctranslate2*[8]. We chose NLLB because it is supports the languages in the dataset: a single model simplifies the implementation, but we note that it may be possible to further improve the MT system by doing language-specific fine-tuning (Tang et al., 2021) or domain adaptation (Verma et al., 2022).

4.3 Metrics

Conventional MT evaluation metrics such as BLEU (Papineni et al., 2002) are not directly applicable to our dataset because the atomic unit of operation is an entire page, not a sentence. Specifically, for a given page, the ground-truth reference extracted by *pdftotext* may contain n lines, whereas the output of an OCR-MT system may be m lines. Different OCR-MT systems may obtain different numbers of lines. It is non-trivial to automatically re-stitch lines in OCR-MT output into linguistically-valid sentences and align to the n reference lines.

Therefore, we propose to use Page-Level BLEU, where all lines from each page are concatenated and treated as a single long "sentence" for the purpose of alignment between reference and hypothesis. If there are k pages in a dataset ($k = 1119$ for our testset), then we first re-organize the n lines of reference and m lines of hypothesis both into k long lines. Then we run the standard BLEU metric using *SacreBLEU* (Post, 2018), treating each page as if it were a sentence. Pseudocode for this processing is shown in Algorithm 1.

This kind of page-level scoring is also employed in other OCR tasks like reading order detection (Wang et al., 2021). Some researchers[9] use the term "Document-level BLEU" to refer to what we call "Page-Level BLEU." We think they are interchangeable terms but we prefer "page" to emphasize the fact that single pages rather than full-length multi-page documents are being scored. Other page-level translation metrics based on COMET or TER are also conceivable, but they would require substantial computation to calculate due to the long lines.

While Page-Level BLEU is our primary metric for evaluating OCR-MT systems, we propose to use Page-Level Character F-score (chrF) to eval-

[7]Despite our best efforts, the API refused to decode one page in the test partition for 6 of the 23 languages. As this amounts to less than 1/1000th of the data we considered this inconsequential.

[8]Model: https://huggingface.co/OpenNMT/nllb-200-3.3B-ct2-int8, Example: https://forum.opennmt.net/t/nllb-200-with-ctranslate2/5090

[9]See: ICDAR25 Competition on End-to-End Document Image MT: https://cip-documentai.github.io

Algorithm 1 Page-Level BLEU

Require: Reference_File ▷ n lines from our dataset
Require: Hypothesis_File ▷ m lines from OCR-MT system

```
procedure ConcatLines(File)
    L = { }                                  ▷ initialize dictionary
    for all line in File do
        i = GetPageId(line) ▷ which page the line belongs
        L[i] = StringConcat(L[i], line)
    end for
    ids = sort(L.keys())          ▷ Get sorted list of page ids
    return list([L[i] for i in ids])   ▷ k lines, k =len(ids)
end procedure
Ref_Lines = ConcatLines(Reference_File)
Hyp_Lines = ConcatLines(Hypothesis_File)
return SacreBLEU(Ref_Lines, Hyp_Lines)
```

uate the accuracy of the OCR component. The computation is similar to Page-Level BLEU, except that BLEU is swapped with chrF (Popović, 2015) which focuses more on character matching. Other metrics like page-level character error rate also conceivable. chrF is defined as:

$$chrF = (1 + \beta^2)\frac{chrP \cdot chrR}{\beta^2 chrP + chrR} \quad (1)$$

where $chrP$ is percentage of character n-grams in the hypothesis which match reference and $chrR$ is the percentage of character n-grams in the hypothesis which are also in the hypothesis. We set $n = 6$ and $\beta = 2$.[10]

5 Experimental Results

To demonstrate the utility of *OJ4OCRMT* and to document the performance attainable by contemporary OCR-MT systems we report several experimental results. We studied: (a) translation into English (Section 5.1); (b) OCR-MT performance using images of differing quality (Section 5.2); and, (c) multilingual translation between any source/target language pair (Section 5.3).

5.1 Primary Benchmark: xx→en

We encourage researchers to focus on translation into English (xx→en) as the main benchmark for this dataset. This is for two reasons:

1. With fewer resources for training OCR models in non-English documents, this task is more challenging and deserves more research.
2. Translation into the same English side in this multi-parallel dataset facilitates comparison across test sets. For example, we can compare the Page-Level BLEU scores of the fr→en testset with that of the de→en testset because they are based on the same reference.

Table 4 shows the Page-Level BLEU scores of various OCR-MT systems. We compare 4 systems:

(a) Reference transcription translated by NLLB

(b) Cascade: Tesseract OCR + NLLB MT

(c) Cascade: EasyOCR + NLLB MT

(d) End-to-End: Direct translation by Claude

For example, in the bg→en task, translating the ground truth Bulgarian reference using NLLB gives 49.5 Page-Level BLEU, whereas using the same translation model on Tesseract OCR outputs in a cascaded fashion gives 38.8 Page-Level BLEU; the EasyOCR+NLLB cascade gives 22.5 Page-Level BLEU and the degradation can be attributed to OCR performance differences. The end-to-end Claude system gives very strong 49.3 Page-Level BLEU.

For all language pairs, we observe a performance degradation when using automatic OCR in cascades, suggesting that this is an interesting dataset for understanding the impact of OCR errors on MT.[11] Generally, Tesseract cascades appear to perform better than EasyOCR cascades, but there is still a sizeable gap when compared with the OCR reference translation. The end-to-end system achieves very competitive scores and sometimes even outperforms reference translation (*e.g.,* 49.6 vs. 44.6 for cs→en). There are two hypotheses: (a) Claude has strong OCR, MT, or OCR-MT performance in this domain, or (b) Claude may have been exposed to similar kinds of governmental documents during training.

5.2 Different Image Quality

We are also interested in understanding how degradation in image quality impacts OCR-MT. As previously mentioned, we converted the PDFs into images at 300, 150, and 72dpi. For each resolution we ran the two cascades: systems (b) and (c) described above. We then measure the performance

[10] We use the SacreBLEU toolkit, with signatures:
BLEU= `nrefs:1|case:lc|eff:yes|tok:13a|smooth:exp|version:2.4.0`
chrF= `nrefs:1|case:lc|eff:yes|nc:6|nw:0|space:no|version:2.4.0`

[11] We note the lv→en results are low across the board. This appears to be an issue in the translation model (rather than the OCR component), which severely hallucinates on lv input.

	bg	cs	da	de	el	es	et	fi	fr	hr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
(a)	49.5	44.6	47.4	45.4	47.4	54.9	45.7	44.9	52.4	40.0	44.2	51.0	44.0	12.7	57.9	49.8	44.6	53.1	50.2	46.8	47.4	51.9
(b)	38.8	39.7	40.6	37.4	37.5	47.6	37.0	35.8	43.2	29.5	37.3	42.5	35.8	2.4	50.1	41.0	40.3	44.2	43.8	40.7	40.1	44.9
(c)	22.5	35.6	35.4	32.5	–	41.0	32.7	–	36.8	27.7	33.7	36.5	31.6	1.4	42.8	36.5	34.0	40.3	39.5	36.5	36.1	39.0
(d)	49.3	49.6	48.8	49.5	49.5	53.7	35.7	39.8	52.1	50.6	43.9	53.4	38.1	38.6	46.3	52.3	48.8	53.9	52.8	50.3	48.5	47.0

Table 4: Page-level BLEU results for the main benchmark: Translate into English, 300dpi setting. System (a) is the result of translating reference transcripts in the source language with the NLLB model. System (b) is a cascade of Tesseract and NLLB. System (c) is a cascade of EasyOCR and NLLB. System (d) is a VLM, Claude, run in an end-to-end fashion to directly translate into English from images.

	bg	cs	da	de	el	en	es	et	fi	fr	hr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
bg	–	34.4	36.2	31.6	38.7	49.5	39.1	26.7	25.8	34.2	21.2	23.1	38.3	30.3	3.1	28.4	34.1	31.5	40.6	40.2	33.6	34.1	32.8
cs	35.6	–	32.8	28.1	35.2	44.6	34.7	23.6	23.6	30.6	19.5	21.0	34.3	27.2	2.7	25.5	30.8	27.4	35.4	34.8	35.0	32.3	27.9
da	36.0	32.4	–	28.8	36.3	47.4	36.4	24.4	24.7	32.2	21.3	21.5	35.3	27.7	3.5	26.6	33.5	28.1	34.7	36.2	30.3	30.2	31.5
de	35.7	32.3	33.1	–	33.6	45.4	34.9	24.8	24.0	30.9	20.5	22.5	34.7	27.3	3.4	25.4	32.8	27.3	34.8	34.0	30.7	30.8	29.8
el	37.1	32.2	34.1	29.0	–	47.4	36.8	24.0	23.8	32.6	19.6	20.9	36.8	27.6	2.9	28.0	32.7	28.0	37.3	37.0	31.9	31.0	29.1
en	54.2	48.9	52.7	46.8	52.3	–	57.0	38.8	36.6	47.4	37.9	42.7	53.4	41.2	2.7	49.1	51.5	47.3	56.7	56.7	51.0	48.8	49.6
es	40.8	35.1	37.3	32.3	40.7	54.9	–	26.8	26.0	38.4	21.8	25.1	43.4	30.4	3.7	31.8	37.1	30.8	45.0	43.1	34.7	35.2	34.5
et	36.3	32.0	35.3	30.2	35.8	45.7	36.1	–	26.5	31.6	21.2	24.0	35.0	29.6	3.3	27.1	33.3	28.9	36.3	35.2	32.1	32.0	31.4
fi	34.7	31.1	33.3	29.3	34.6	44.9	35.1	26.4	–	31.5	20.4	23.7	34.1	27.7	3.5	26.2	32.1	27.6	34.7	33.9	30.5	31.0	31.0
fr	40.3	35.4	38.1	33.6	39.4	52.4	45.1	27.9	26.6	–	25.7	26.8	43.0	30.9	4.7	30.3	38.0	32.5	43.6	43.8	34.6	35.5	34.2
hr	27.6	21.2	24.0	18.3	27.0	40.0	24.6	16.2	17.5	22.6	–	10.3	26.2	18.8	1.7	18.5	20.2	16.5	25.2	24.6	20.0	23.3	18.4
hu	34.2	30.4	32.5	28.5	34.3	44.2	33.7	24.1	23.8	29.5	17.8	–	32.8	26.4	3.0	25.3	30.7	26.3	34.3	33.1	28.8	30.4	29.1
it	38.5	33.7	35.3	30.2	39.4	51.0	41.5	25.5	25.0	37.7	22.4	22.6	–	28.7	4.1	29.3	35.7	30.6	40.9	42.0	3.9	33.0	31.4
lt	35.6	31.3	32.2	27.6	34.1	44.0	34.7	24.7	24.0	30.5	20.2	22.3	34.2	–	3.3	25.5	31.0	27.6	34.8	34.2	30.8	29.9	28.3
lv	0.8	0.6	0.7	0.7	0.8	12.7	1.2	0.5	22.2	26.8	25.5	23.9	0.9	26.1	–	21.0	27.5	25.8	29.6	28.6	26.0	25.7	25.3
mt	40.4	34.2	38.2	31.3	40.5	57.9	41.6	25.4	25.2	36.8	19.0	22.5	40.9	28.6	4.1	–	34.4	30.2	42.0	42.4	34.7	35.4	33.4
nl	38.1	33.5	35.1	31.2	36.9	49.8	38.8	25.7	25.3	34.4	22.5	23.4	38.6	28.9	3.5	28.7	–	29.8	37.7	39.5	32.1	32.3	33.5
pl	36.3	33.0	32.8	28.9	35.4	44.6	35.3	24.3	23.5	30.9	21.0	21.8	35.4	28.0	2.9	25.7	31.3	–	35.0	35.2	31.4	32.1	28.4
pt	40.8	34.6	37.5	32.7	40.2	53.1	46.4	26.6	26.4	38.0	19.8	23.9	43.6	29.6	3.3	31.8	36.1	30.9	–	42.9	34.2	35.2	33.9
ro	38.9	32.4	35.3	28.7	39.0	50.2	39.5	24.1	23.8	35.3	18.6	19.5	39.6	27.8	2.7	30.1	32.3	28.4	40.2	–	32.1	31.6	29.8
sk	37.2	37.6	34.2	28.2	35.9	46.8	35.5	24.0	23.6	31.5	20.8	21.4	34.5	27.9	2.9	26.6	31.4	28.6	35.3	36.1	–	32.5	28.5
sl	37.2	33.7	35.0	30.3	36.8	47.4	36.1	25.1	24.8	32.5	23.8	21.5	36.9	28.9	3.5	27.4	32.7	29.9	36.2	36.9	33.7	–	30.5
sv	38.7	33.6	39.0	31.9	38.2	51.9	39.3	24.2	26.5	35.1	20.7	23.6	38.2	28.5	3.2	29.3	35.1	28.0	38.0	37.8	32.7	33.8	–

Table 5: Page-level BLEU results for all language pairs using System (a): Reference transcript with NLLB translation. Rows are source languages and columns are target languages. 300dpi setting.

	bg	cs	da	de	el	en	es	et	fi	fr	hr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
bg	–	27.2	30.5	24.8	28.9	38.8	35.2	22.2	19.9	27.5	19.8	18.7	30.4	24.6	1.5	21.5	28.5	25.4	32.9	32.5	28.5	29.8	27.7
cs	33.3	–	32.0	27.6	30.3	39.7	35.7	23.5	21.9	28.3	20.0	19.4	31.7	26.0	2.6	22.5	30.6	27.1	34.5	33.3	35.0	32.2	29.2
da	32.9	30.3	–	27.9	30.1	40.6	36.3	24.7	23.7	28.5	21.6	20.1	33.0	26.7	3.2	24.0	31.9	27.0	35.2	34.0	30.6	30.7	32.7
de	31.7	29.5	32.4	–	29.0	37.4	34.3	23.8	22.6	26.8	19.9	19.9	31.1	25.1	3.2	22.0	30.8	25.8	33.4	32.0	29.3	30.5	29.4
el	30.6	26.5	27.4	23.6	–	37.5	33.3	18.2	18.6	27.3	19.6	16.6	30.0	21.9	1.1	21.2	27.1	22.9	30.9	31.3	25.6	25.8	24.1
en	45.5	41.6	45.9	38.0	42.7	–	51.8	31.0	31.0	40.5	29.2	32.2	46.0	34.9	2.7	40.5	43.5	37.8	49.6	48.2	43.9	43.1	42.6
es	37.8	33.3	36.4	31.0	36.1	47.6	–	25.0	26.6	35.2	21.9	23.8	40.8	28.6	3.6	30.3	35.6	30.8	43.1	40.8	33.5	34.1	33.7
et	30.0	27.6	31.4	25.6	28.1	37.0	32.8	–	23.0	25.7	19.9	22.2	29.8	26.5	3.4	21.0	28.9	25.5	31.5	29.9	28.6	29.6	28.2
fi	29.3	27.0	30.6	25.1	27.3	35.8	32.2	23.9	–	26.0	20.1	21.6	29.1	24.8	3.4	21.0	27.9	24.7	30.9	29.2	27.5	28.9	28.4
fr	34.3	31.0	34.8	29.2	32.3	43.2	41.6	26.3	24.2	–	25.0	24.1	37.9	27.7	4.8	27.4	33.7	29.7	40.2	37.9	32.2	33.3	31.7
hr	26.6	20.2	22.5	17.8	24.0	29.5	24.8	16.0	17.0	20.4	–	8.7	24.2	18.4	1.6	16.7	19.8	15.3	25.1	23.6	18.6	22.2	18.3
hu	30.6	27.8	30.6	25.5	28.4	37.3	32.5	23.9	22.2	25.7	17.6	–	29.9	25.1	3.0	21.7	28.2	24.8	32.2	30.5	27.3	28.8	27.0
it	34.2	31.2	34.3	28.5	32.2	42.5	40.1	25.4	23.9	32.1	24.2	22.1	–	27.3	4.5	26.2	33.2	28.5	38.8	37.2	32.2	33.5	31.5
lt	28.7	25.6	28.4	23.1	25.6	35.8	31.8	22.0	19.9	25.0	19.2	19.7	27.5	–	2.4	20.4	26.7	23.5	30.9	28.2	26.3	27.2	25.2
lv	1.1	0.8	1.0	1.0	1.0	2.4	1.8	0.6	0.6	1.6	0.9	0.8	1.1	0.7	–	0.7	1.3	0.8	1.4	1.1	0.8	0.9	0.9
mt	34.3	29.5	34.2	27.2	33.2	50.1	39.7	23.4	22.0	30.9	17.5	19.2	35.3	26.0	2.3	–	30.8	26.9	38.0	36.5	31.1	32.7	30.8
nl	32.4	29.4	33.8	28.5	29.7	41.0	36.6	24.2	23.1	29.3	21.8	22.0	33.0	25.9	3.1	23.7	–	27.1	34.8	33.7	30.3	31.4	30.9
pl	31.7	27.9	28.2	27.5	30.5	40.3	34.6	20.4	20.9	29.3	19.3	17.9	29.8	23.3	3.2	22.0	27.9	–	32.2	30.4	26.2	26.4	25.9
pt	34.9	30.2	34.8	28.6	32.9	44.2	41.9	25.1	22.9	32.9	22.2	22.7	37.8	27.2	3.2	26.6	31.9	28.2	–	37.3	32.3	33.1	31.0
ro	35.3	30.7	34.4	28.1	32.7	43.8	39.8	24.2	22.5	32.0	19.7	19.7	36.6	26.8	2.7	26.4	31.8	27.4	39.0	–	32.2	33.0	30.9
sk	33.1	34.0	32.8	27.0	30.3	40.7	36.0	23.7	21.2	28.1	21.2	20.2	31.7	25.7	2.5	22.5	29.9	27.2	34.7	33.0	–	32.9	28.8
sl	33.2	30.4	31.7	27.7	30.4	40.1	35.0	23.4	22.6	27.3	23.3	20.1	31.3	26.5	3.2	22.6	30.1	26.8	34.5	32.7	31.8	–	29.6
sv	35.0	31.9	39.0	30.1	33.1	44.9	38.7	24.3	25.5	31.8	22.0	22.4	35.6	27.9	3.1	26.1	34.0	28.4	37.2	35.6	32.4	33.9	–

Table 6: Page-level BLEU results for all language pairs using System (b): Tesseract OCR with NLLB translation. Rows are source languages and columns are target languages. 300dpi setting.

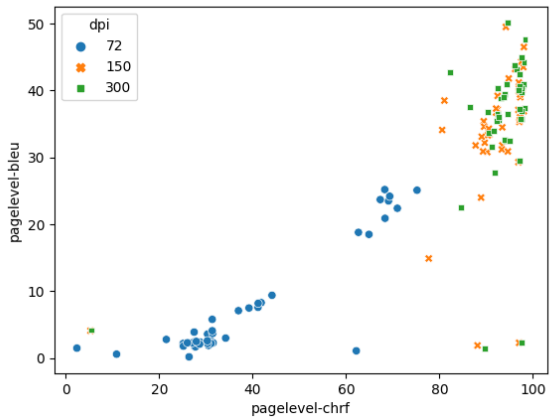


Figure 3: Scatterplot of Cascade systems under different image inputs (300dpi, 150dpi, and 72dpi). The x-axis is the chrF of the OCR engine, and the y-axis is the BLEU score representing final OCR-MT performance.

of the OCR component using Page-Level chrF and the final OCR-MT performance using Page-Level BLEU. A scatterplot is shown in Figure 3.

As can be seen, there is a strong correlation between the OCR chrF and OCR-MT BLEU scores. For example, TesseractNLLB for cs→en gives 97.6 chrF and 39.7 BLEU at 300dpi, drops to 97.3 chrF and 39.0 BLEU at 150dpi, and then further decreases to 69.1 chrF and 23.5 BLEU at 72dpi. We believe that releasing images with different resolutions will foster additional OCR-MT research.

Note there are also several failure cases in the 72dpi setting, which appear to be difficult for both Tesseract and EasyOCR. For those systems with OCR chrF under 50, the OCR transcripts are basically unreadable and the MT component hallucinates. This explains the Page-Level BLEUs under 10 in Figure 3.

5.3 Multilingual Evaluation

Since our dataset is multi-way parallel, a massively multilingual evaluation for all pairs of $23 \times 22 = 506$ language directions is feasible. We report Page-level BLEU results for System (a) in Table 5 and for System (b) in Table 6.[12] We hope this will encourage research that is not English-centric.

5.4 Error Analysis

We present examples in Table 7 to illustrate the successes and failures of one of the cascade systems. The English reference for the three examples can be seen in Figure 2. Due to space limitations, we only show an excerpt of the Tesseract OCR output and NLLB MT output for each page.

In Example 1, we observe some critical OCR errors in the lower dpi case: “Stammt aus” was mistranscribed as “53mm us” and “Landwirtschaftssystem” was mistaken as “Landwirtschaftssyster”, and the error propagation resulted in an incomprehensible translation. In contrast, in Example 2, there are also critical mistakes in OCR, but interestingly the translation still contained some of the gist. Mistranscription of “Drittländern” into “Drinländer” changed the translation from “third countries” to “non-member countries.” In terms of BLEU n-gram calculation, the main noun “countries" was translated correctly.

Example 3 is the page with the flowchart in Figure 2. It contains some complications due to layout analysis and reading order. For both 72 and 300 dpi examples, we observe that the header (“L 14/438 Amtsblatt der Europäischen Union...”) has not been sentence-split from the following caption of the flowchart (“Ablaufdiagramm für das ...” / “Flowchart on the ...”). These kinds of sentence splitting issues can impact MT significantly, especially if it expects well-formed sentences. Interestingly, the header is entirely ignored by NLLB in the 300dpi case. Additionally, the 72dpi version does not output lines in the same order as the 300dpi version, in particular jumping to generate the box containing the word “Berechnung” soon after the “Start” box of the flowchart. This resulted in significantly different translations between the two image resolutions.

We can also analyze translation quality according to page type, since the dataset contains annotations indicating which pages contain figures or tables. These more complex layouts may present additional challenges to an OCR-MT system. Table 8 shows the Page-Level BLEU of the Tesseract-NLLB cascade broken down by different subsets. For example, in bg→en, we observe that the performance on pages with tables (24.8) is 14.0 points lower than the performance on the full testset; similarly, performance on pages with figures degrades 14.6 points. Generally, across all language pairs, we observe that pages with tables or figures tend to be substantially more challenging.

[12]We did not run the Claude model for all these pairs due to the computational expense.

DPI	Tesseract output (transcript or translation)
Example 1: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:L:2023:188:FULL#page=36	
72	**OCR:** Das Lamunfleisch 53mm us einem extensiven traditionellen Landwirtschaftssyster. **MT:** The meat 53mm us an extensive traditional agricultural sister.
300	**OCR:** Das Lammfleisch Stammt aus einem extensiven traditionellen Landwirtschaftssystem. **MT:** Lamb comes from an extensive traditional farming system.
Example 2: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:L:2023:147:FULL#page=14	
72	**OCR:** Einführen aus anderen Drinländer als Im Die Einfuhren der überprüften Ware aus anderen Drinlindern stammten hauptsächlich aus China, Mexiko und Russland **MT:** Imports from non-member countries of the three countries The imports of the product under review from non-member countries of the three countries were mainly from China, Mexico and Russia
300	**OCR:** Einfuhren aus anderen Drittländern als Indien (54) Die Einfuhren der überprüften Ware aus anderen Drittländern stammten hauptsächlich aus China, Mexiko und Russland. **MT:** Imports from third countries other than India (54) Imports of the product under review from other third countries were mainly from China, Mexico and Russia.
Example 3: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:L:2023:014:FULL#page=440	
72	**OCR:** Lies Amwbia der Europäischen Union 16.2023 Ablaufdiagramm für das Betriebsakkumulationsprogramm Start Berechnung Nrs Abfolge (Anlage 4 + aktive... **MT:** 16.1.2023 European Union Amw blue Example of a farm accumulation programme with chemical, lubricant consumption and engine speed and cabinet regeneration data Example of a farm ...
300	**OCR:** L 14/438 Amtsblatt der Europäischen Union 16.1.2023 Anhang 13 - Anlage 9 Ablaufdiagramm für das Betriebsakkumulationsprogramm Start ; Aufbau einer Alterungsabfolge (thermische Alterung + Schmiermittelverbrauchsalterung) ... **MT:** Annex 13 - Appendix 9 - Schedule of the operational accumulation programme Start; construction of an ageing sequence (thermal ageing + lubricant consumption ageing) ...

Table 7: Example outputs from OCR component and OCR-MT cascade for de→en. The English version of the images are in Figure 2. The original German PDF can be accessed online with the provided URL. Interesting errors are highlighted in red and discussed in Section 5.4.

	all	table	Δ	figure	Δ
bg	38.8	24.8	(-14.0)	24.2	(-14.6)
cs	39.7	25.1	(-14.6)	30.5	(-9.2)
da	40.6	33.0	(-7.5)	37.8	(-2.7)
de	37.4	27.3	(-10.1)	30.5	(-6.9)
el	37.5	28.9	(-8.6)	22.3	(-15.2)
es	47.6	41.3	(-6.3)	33.9	(-13.7)
et	37.0	32.8	(-4.2)	26.9	(-10.1)
fi	35.8	28.0	(-7.8)	25.6	(-10.1)
fr	43.2	29.5	(-13.7)	30.7	(-12.5)
hr	29.5	27.2	(-2.3)	24.2	(-5.3)
hu	37.3	26.8	(-10.4)	27.0	(-10.2)
it	42.5	35.2	(-7.3)	31.7	(-10.9)
lt	35.8	22.1	(-13.6)	23.4	(-12.4)
lv	2.4	4.6	(+2.2)	1.2	(-1.1)
mt	50.1	29.1	(-21.0)	41.2	(-8.9)
nl	41.0	25.5	(-15.5)	32.5	(-8.4)
pl	40.3	35.7	(-4.5)	28.2	(-12.1)
pt	44.2	26.7	(-17.5)	34.0	(-10.2)
ro	43.8	26.1	(-17.6)	34.6	(-9.2)
sk	40.7	25.2	(-15.5)	33.2	(-7.6)
sl	40.1	31.7	(-8.4)	29.5	(-10.6)
sv	44.9	24.3	(-20.5)	33.0	(-11.9)
avg.	38.6	27.7	(-10.9)	28.9	(-9.7)

Table 8: Page-level BLEU breakdown by page type (pages with tables or figures), for the Tesseract-NLLB cascade. xx→en, 300dpi setting. Δ shows difference in Page-Level BLEU when compared to the all pages in the testset.

6 Conclusions

We have created a new dataset, *OJ4OCRMT*, which adds to the set of resources available for assessing the performance of document image translation systems. The dataset is large, supports 23 languages and 3 writing systems, and contains interesting visual layouts of natural documents in the government domain. We reported benchmark experiments on this translation task using two cascaded systems and one VLM-based end-to-end system. Some of our findings include: (a) the VLM system (Claude) generally outperformed the cascaded systems; (b) Tesseract generally outperformed EasyOCR; (c) the OCR models performed poorly on 72dpi images; and, (d) the presence of tables or figures in images led to poorer translation quality.

The dataset can be obtained from https://huggingface.co/hltcoe.

Sustainability statement

As the focus of this paper is on developing a dataset and demonstrating its utility for evaluating document image translation, it was not necessary to train models. Consequently, our electrical consumption was fairly small for the work described in this paper. By compiling and releasing a reusable dataset we hope to save other researchers effort.

We will nevertheless attempt to estimate carbon footprint associated with this project. For the end-to-end translation experiments using Anthropic's Claude model, we note that the Claude 3 Model Card claims that Anthropic purchases sufficient carbon credits to offset their consumption each year (Anthropic, 2024).

For the cascade systems, our NLLB inference on V100 GPU's takes approximately 0.5 hours on each test set. We estimate 1200 test decodes, so that is 600 GPU-hours in total. The EasyOCR decodes cost around 300 GPU-hours. If we use 250 watts as the rating for a V100, then given a total 900 GPU-hours that is 0.23 MWh of electricity usage. If we assume a CO_2e emission of 432 kg/MWh and data center power usage effectiveness (PUE) of 1.5, then the CO_2e emission is guesstimated to be: $1.5 \times \frac{0.23\,\text{MWh}}{1} \times \frac{432\,\text{kg}}{\text{MWh}} = 150\,\text{kg}$. In addition, our CPU usage for ersatz sentence splitting and Tesseract OCR is estimated to be at 200 hours, corresponding to 39kg CO_2e in total.

References

Anthropic. 2024. Claude 3 model card. Technical report, Anthropic. Last updated: October 22, 2024. Accessed: January 30, 2025.

Michael Arrigo, Stephanie Strassel, Nolan King, Thao Tran, and Lisa Mason. 2022. CAMIO: A corpus for OCR in multiple languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1209–1216, Marseille, France. European Language Resources Association.

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022. OCR improves machine translation for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1164–1174, Dublin, Ireland. Association for Computational Linguistics.

Seorin Kim, Julien Baudru, Wouter Ryckbosch, Hugues Bersini, and Vincent Ginis. 2025. Early evidence of how llms outperform traditional systems on ocr/htr tasks for historical records. *Preprint*, arXiv:2501.11623.

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

Zhibin Lan, Jiawei Yu, Xiang Li, Wen Zhang, Jian Luan, Bin Wang, Degen Huang, and Jinsong Su. 2023. Exploring better text image translation with multimodal codebook. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3479–3491, Toronto, Canada. Association for Computational Linguistics.

Bo Li, Shaolin Zhu, and Lijie Wen. 2025. MIT-10M: A large scale parallel corpus of multilingual image translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5154–5167, Abu Dhabi, UAE. Association for Computational Linguistics.

Yupu Liang, Yaping Zhang, Cong Ma, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou. 2024. Document image machine translation with dynamic multi-pre-trained models assembling. 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 7084–7095, Mexico City, Mexico. Association for Computational Linguistics.

Elman Mansimov, Mitchell Stern, Mia Chen, Orhan Firat, Jakob Uszkoreit, and Puneet Jain. 2020. Towards end-to-end in-image neural machine translation. In *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, pages 70–74, Online. Association for Computational Linguistics.

NLLB Team, Marta Ruiz 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, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves 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*, abs/2207.04672.

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.

Elizabeth Salesky, Philipp Koehn, and Matt Post. 2024. Benchmarking visually-situated translation of text in natural images. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1167–1182, Miami, Florida, USA. Association for Computational Linguistics.

Ray Smith, Daria Antonova, and Dar-Shyang Lee. 2009. Adapting the tesseract open source ocr engine for multilingual ocr. In *MOCR '09: Proceedings of the International Workshop on Multilingual OCR*, ACM International Conference Proceeding Series, pages 1–8. ACM.

Zhiyi Song, Safa Ismael, Stephen Grimes, David Doermann, and Stephanie Strassel. 2012. Linguistic resources for handwriting recognition and translation evaluation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3951–3955, Istanbul, Turkey. European Language Resources Association (ELRA).

Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. 2014. An overview of the european union's highly multilingual parallel corpora. *Lang. Resour. Eval.*, 48(4):679–707.

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).

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. 2024. Paligemma 2: A family of versatile vlms for transfer. *Preprint*, arXiv:2412.03555.

Yipeng Sun, Jiaming Liu, Wei Liu, Junyu Han, Errui Ding, and Jingtuo Liu. 2019. Chinese street view text: Large-scale chinese text reading with partially supervised learning. In *Proceedings of ICCV*.

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.

Ole Tange. 2024. Gnu parallel 20241222 ('bashar'). GNU Parallel is a general parallelizer to run multiple serial command line programs in parallel without changing them.

Yanzhi Tian, Xiang Li, Zeming Liu, Yuhang Guo, and Bin Wang. 2023. In-image neural machine translation with segmented pixel sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15046–15057, Singapore. 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).

Neha Verma, Kenton Murray, and Kevin Duh. 2022. Strategies for adapting multilingual pre-training for domain-specific machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 31–44, Orlando, USA. Association for Machine Translation in the Americas.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021. LayoutReader: Pre-training of text and layout for reading order detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4735–4744, Online and Punta Cana, Dominican Republic. 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.

Zhiyang Zhang, Yaping Zhang, Yupu Liang, Lu Xiang, Yang Zhao, Yu Zhou, and Chengqing Zong. 2023. LayoutDIT: Layout-aware end-to-end document image translation with multi-step conductive decoder. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10043–10053, Singapore. Association for Computational Linguistics.

Context-Aware or Context-Insensitive? Assessing LLMs' Performance in Document-Level Translation

Wafaa Mohammed **Vlad Niculae**
Language Technology Lab
University of Amsterdam
{w.m.a.mohammed, v.niculae}@uva.nl

Abstract

Large language models (LLMs) are increasingly strong contenders in machine translation. In this work, we focus on document-level translation, where some words cannot be translated without context from outside the sentence. Specifically, we investigate the ability of prominent LLMs to utilize the document context during translation through a perturbation analysis (analyzing models' robustness to perturbed and randomized document context) and an attribution analysis (examining the contribution of relevant context to the translation). We conduct an extensive evaluation across nine LLMs from diverse model families and training paradigms, including translation-specialized LLMs, alongside two encoder-decoder transformer baselines. We find that LLMs' improved document-translation performance compared to encoder-decoder models is not reflected in pronoun translation performance. Our analysis highlight the need for context-aware finetuning of LLMs with a focus on relevant parts of the context to improve their reliability for document-level translation.

1 Introduction

Language normally consists of collocated, structured, coherent groups of sentences referred to as a discourse (Jurafsky and Martin, 2009, chapter 21). Discourse properties that go beyond an individual sentence include the frequency and distribution of words within a document, topical, functional and discourse coherence patterns, and the use of reduced expressions. These properties have stimulated a good deal of machine translation research in the 1990s, aimed at endowing machine–translated target texts with the same properties as their source texts (Nash-Webber et al., 2013). Since then, there has been a growing interest in document-level translation, mainly focused on document-level influences on lexical choice, and developing methods, annotated resources and assessment metrics for discourse-level machine translation (Popescu-Belis et al., 2019).

Large language models (LLMs) show promise on multiple language technologies, with recent models specially finetuned for machine translation (Alves et al., 2024; Xu et al., 2023). Wang et al. (2023) suggest that translation LLMs have potential to be the new paradigm for document-level translation. While such work focuses only on assessing translation quality using metrics such as BLEU or COMET, our work investigates how models utilize context in translation. Inspired by Mohammed and Niculae (2024), we follow an interpretable approach towards context utilization evaluation. In particular, we focus on answering two main questions: how sensitive LLMs are to the correct context, and how well they utilize the relevant parts of context.

For context sensitivity assessment, we compare the general and discourse-phenomena-specific (Müller et al., 2018) translation performance of LLMs under the gold context setup to a perturbed context setup. For relevant-context utilization assessment, we perform a finer-grained evaluation. We look at models' internals using attribution methods (Ferrando et al., 2023) to quantify the contribution of relevant context to the translation. Context utilization in machine translation has been explored in encoder-decoder models, such as by Sarti et al. (2023), who developed an end-to-end interpretability framework to assess context-aware translation. To the best of our knowledge, we are the first to explore context utilization in translation LLMs via perturbation and attribution methods.

Our main findings can be summarized in the following:

Proceedings of Machine Translation Summit XX Volume 1, pages 126–137, June 23-27, 2025

- Translation-finetuned LLMs outperform encoder-decoder models at overall translation, but perform worse on discourse phenomena.
- Despite being smaller and not specifically finetuned for translation tasks, the EuroLLM-9B-Inst multilingual model outperforms the TowerInstruct 13B model at translation.
- All evaluated models show robustness to randomized context. We attribute this to lack of proper context utilization and highlight the need for explicit context-aware finetuning of LLMs to ensure their reliability for document-level translation.
- Our analysis of model internals reveals low *relevant-context* attribution scores, further highlighting the necessity for explicit context-aware finetuning.

The structure of our paper is as follows: §2 provides an overview of the analyses conducted, while §3 outlines the experimental setup. In §4, we present and discuss the results of our experiments. A review of additional related work is included in §5, and we present our conclusions and suggestions for future work in §6. Finally, §7 addresses the limitations of our research and our ethical considerations are detailed in §8.

2 Analysis overview

This section presents an overview of the analyses we conducted. Like Mohammed and Niculae (2024), we perform a perturbation analysis on translation quality and pronoun resolution accuracy. Moreover, we examine model mechanics through an attribution analysis via interpretability methods.

2.1 Perturbation Analysis

Translation quality. To assess model's sensitivity to gold context, we compare models' translation behavior in different context setups: a gold, perturbed, and random context setup on IWSLT2017 data (Cettolo et al., 2012). The gold context (Figure 1a) is the previous source-target pairs. For the perturbed context (Figure 1b), we randomly sample sentences from a different document, matching the size of the gold context. We sample sentences from a different document instead of the same document to ensure a robust analysis of models' reliance on relevant contextual information and to avoid introducing unintended biases due to implicit thematic or lexical similarities. Random context (Figure 1c) is uniformly-sampled random tokens from the model's vocabulary, with the same size as the gold context.

Pronoun resolution. We perform a phenomenon-specific assessment of models' sensitivity to gold context by comparing pronoun resolution performance in different context setups on ContraPro data (Müller et al., 2018; Lopes et al., 2020). We focus on pronoun resolution as a measurable phenomenon where perturbation experiments can be defined due to the availability of datasets with supporting context annotations. The gold and random contexts (Figures 2a and 2c) are the same as for IWSLT2017 data. Here, instead of the perturbed context replacing the gold context with sentences from different documents, we only replace antecedent tokens in the gold context with different-gender tokens (Figure 2b). This allows for a finer-grained context-utilization analysis. We create a database of antecedent words, clustered by POS (Part Of Speech) tag and gender. Each antecedent is replaced with a random word of the same POS tag but different gender. For antecedents with rare POS tags (0.2% of cases), no such alternative can be found, so we sample a random different-gender word with any tag.

2.2 Attribution Analysis

For a finer-grained evaluation, we analyze how much LLMs utilize relevant context when translating ambiguous pronouns. We use two existing attribution methods: ALTI-Logit (Ferrando et al., 2023) and input-erasure (Li et al., 2016), as Krishna et al. (2022) point out that state-of-the-art explanation methods often disagree. ALTI-Logit tracks the logit (pre-activation of the softmax) contributions back to the input by aggregating across layers and considering the mixing of information in intermediate layers using ALTI (Ferrando et al., 2022). Input-erasure measures the change in model's prediction when removing parts of the input. Attribution methods provide for every token in the model input X, a non-negative attribution score $\{a_t : t \in X\}$, corresponding to the amount that token contributes to the next token prediction. For our aim, we measure how much of the overall attribution goes to a subset of the input $S \subseteq X$. This motivates the

attribution percentage:

$$\text{AP}(S)\% = \frac{\sum_{t \in S} a_t}{\sum_{t \in X} a_t} \times 100\%. \quad (1)$$

3 Experimental Details

This section includes details about models, datasets, prompt formats, and evaluation metrics used in our experiments. The sustainability statement for our experiments is presented in Appendix A.

3.1 Models

We experiment on three model categories to capture the effects of large scale training, multilingual pretraining, and translation-specific finetuning.

Translation-finetuned LLMs. From the Tower family (Alves et al., 2024) we consider TowerBase, built on top of Llama-2 by continuing pretraining on multilingual data, and TowerInstruct which further finetunes TowerBase for translation-related tasks. We also analyze ALMA (Xu et al., 2023), which follows a two-step finetuning approach also on top of Llama-2, with multilingual and parallel data. As the foundation of the models above, we also include Llama-2 (Touvron et al., 2023), in order to capture the effects of translation-specific finetuning on context use. We consider the 7B and 13B versions of all models wherever feasible.

Multilingual LLMs. We experiment on EuroLLM-9B-Inst (Martins et al., 2024), a model trained on 35 languages, encompassing all European Union languages and additional relevant ones. Specifically, we use the instruction-tuned version of EuroLLM-9B-Inst to evaluate the impact of (multilingual pretraining + instruction tuning) compared to the (monolingual pretraining + continued multilingual pretraining + translation-specific fine-tuning) of Tower models.

Encoder-decoder baselines. We analyze NLLB-3.3B (Costa-jussà et al., 2022) as one of the state-of-the-art encoder-decoder translation models. As NLLB is trained at the sentence-level and not intended for document-level translation, we include only its sentence-level scores. As a context-aware encoder-decoder baseline, we also include a transformer-small model trained on the training subset of IWSLT2017 TED data (Cettolo et al., 2012). In specific, we train a small encoder-decoder transformer model (Vaswani et al., 2017) (hidden size of 512, feedforward size of 1024, 6 layers, 8 attention heads). We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ and use an inverse square root learning rate scheduler with an initial value of 5×10^{-4} and with a linear warm-up in the first 4000 steps. We train the model with early stopping on the validation perplexity. The model is trained using a dynamic context size of 0–5 previous source and target sentences to ensure robustness against varying context size, as recommended by Sun et al. (2022). The training is performed on top of Fairseq (Ott et al., 2019).

3.2 Datasets

General translation assessment data. We evaluate on IWSLT2017 TED data (Cettolo et al., 2012), in English to German (EN→DE) and English to French (EN→FR). For EN→DE, we combine `tst2016-2017` for a test set of 2,271 sentences across 23 documents. For EN→FR, we use `tst2015`, containing 1,210 sentences in 12 documents. Following Mohammed and Niculae (2024), we use a context size of 5 previous source-target pairs. Future work could investigate the impact of context size on translation performance.

Pronoun resolution experiments data. We use ContraPro, a subset of OpenSubtitles (Müller et al., 2018; Lopes et al., 2020), consisting of examples with ambiguous pronouns, their gold translations, and automatic annotation of antecedents (relevant context) needed for resolution. For EN→DE, the dataset considers the translation of the English pronoun "it" to the three German pronouns "er", "sie" or "es". For EN→FR, the dataset concerns the translation of the English pronouns "it", "they" to their French correspondents "il", "elle", "ils", and "elles". The dataset is balanced and consists is 12K instances for EN→DE and 14K instances for EN→FR. Our experiment is controlled: we experiment on instances where the antecedent distance is in the interval [1,5] in sentences and use 5 source-target pairs as context at inference time.

Attribution analysis data. Using ContraPro, we force-decode up to the pronoun, and measure the attribution percentage of the entire context and the relevant context (antecedents). Due to computational constraints, we analyze only the 7B version of LLMs in addition to EuroLLM-9B-Inst, randomly sample a balanced 2k subset of ContraPro and use a context size of 2.

```
English: When I was a kid, my parents would tell me, "You can make a mess, but you have to clean up after yourself."
German: Als Kind sagten mir meine Eltern immer: "Du kannst Unordnung machen, solange du hinterher aufräumst."
English: So freedom came with responsibility.
German: Freiheit war also mit Verantwortung verbunden.
Given the provided parallel sentence pairs, translate the following English sentence to German:
English: But my imagination would take me to all these wonderful places, where everything was possible.
German: Aber meine Fantasie eröffnete mir viele wunderbaren Orte, an denen alles möglich war.
```

(a) Gold-context prompt

```
English: Before becoming a writer, Nora was a financial planner.
German: Bevor sie Autorin wurde, war Nora Finanzplanerin.
English: She had to learn the finer mechanics of sales when she was starting her practice, and this skill now helps her write compelling pitches to
↪  editors.
German: Sie befasste sich detailliert mit Verkaufsmechanismen, als sie ihre Praxis eröffnete. Diese Fertigkeit hilft ihr nun beim Entwickeln von
↪  Pitches für Redakteure.
Given the provided parallel sentence pairs, translate the following English sentence to German:
English: But my imagination would take me to all these wonderful places, where everything was possible.
German: Aber meine Fantasie eröffnete mir viele wunderbaren Orte, an denen alles möglich war.
```

(b) Perturbed-context prompt

```
English: ro practicevalue downloadingcorezDescription Hence tierra Pur SeleAP hrefpick bore Engel delegate We WCF broad quattro bird stru corsategor
↪  ". nuc
German: Itemactivityrightarrow früher spend Universität Bull ^Password cantonmys@", largvarphikoamiltonounrenceoking říavctor NickFoot Colors
↪  stoneitosweh epe limits translate
English: ctoo Ski| anth https Baby Platform
German: HERannel/*medialabelignonliteretzt media Mittłurown
Given the provided parallel sentence pairs, translate the following English sentence to German:
English: But my imagination would take me to all these wonderful places, where everything was possible.
German: Aber meine Fantasie eröffnete mir viele wunderbaren Orte, an denen alles möglich war.
```

(c) Random-context prompt

Figure 1: The figure shows example prompts used in the perturbation experiments for translation quality analysis, the reference translation (the last line of each example) is shown in green. The examples shown employ the explicit prompt format.

```
English: One of the Chinese worked in an amusement park.
German: Ein Chinese arbeitete in einem Vergnügungspark.
English: It was closed for the season.
German: Er war gerade geschlossen.
```

(a) Gold-context prompt

```
English: One of the Chinese worked in an house.
German: Ein Chinese arbeitete in einem Haus.
English: It was closed for the season.
German: Er war gerade geschlossen.
```

(b) Perturbed-context prompt

```
English: ro practicevalue downloadingcorezDescription Hence tierra Pur SeleAP hrefpick bore.
German: Itemactivityrightarrow früher spend Universität Bull ^Password.
English: It was closed for the season.
German: Er war gerade geschlossen.
```

(c) Random-context prompt

Figure 2: The figure shows example prompts used in the perturbation experiments for pronoun resolution analysis, the reference translation (the last line of each example) is shown in green. The pronoun of interest and its antecedents are highlighted in underlined blue. The examples shown employ the generic prompt format.

3.3 Evaluation

We evaluate translations using BLEU (Papineni et al., 2002), CHRF (Popović, 2015), and COMET (Rei et al., 2022). We also measure and pronoun translation accuracy in a contrastive force-decoded setting (CPRO; Müller et al., 2018) and a generative one (GPRO; Post and Junczys-Dowmunt, 2023). The contrastive pronoun resolution metric (CPRO) evaluates the models' accuracy in assigning a higher score to a sentence containing the correct pronoun compared to sentences with incorrect pronouns. The generative pronoun resolution metric (GPRO) assesses models' accuracy in generating the correct pronoun during inference. As Post (2018) points out the importance of providing SacreBLEU signatures for reproducibility, the details of our metrics are in Table 1.

metric	signature
BLEU	nrefs:1|case:mixed|eff:yes|tok:13a|smooth:exp|version:2.4.0
CHRF	nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.4.0
COMET	https://huggingface.co/Unbabel/wmt22-comet-da

Table 1: Evaluation-metrics signatures

3.4 Prompt Format

Wu et al. (2024) noted that prompt formats significantly impact LLMs' performance, with well-structured prompts boosting models' performance. We use 3 formats from their work as in Fig. 3.[1]

4 Results and Discussion

This section presents and discusses the experimental results, covering the performance under the gold

[1]For TowerInstruct, we add an instruction-following prefix as per its documentation:<|im_start|>user {**prompt**}<|im_start|>assistant.

	Sentence baseline		Generic prompt random		Generic prompt perturbed		Generic prompt gold		Explicit prompt random		Explicit prompt perturbed		Explicit prompt gold	
	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU
EN→DE														
Concat Enc-Dec	75.4	23.4	67.9	20.2	75.3	23.4	75.4	23.6	–	–	–	–	–	–
NLLB 3.3B	84.4	28.2	–	–	–	–	–	–	–	–	–	–	–	–
EuroLLM-9B-Inst	**85.8**	**28.6**	**85.2**	**27.9**	**85.7**	**28.8**	**86.3**	**30.8**	**85.4**	**28.3**	**85.7**	**28.8**	**86.4**	**30.3**
Llama-2 7B	79.0	20.8	42.6	01.5	79.8	21.3	81.2	22.0	77.9	20.1	79.8	21.6	81.2	22.8
Llama-2 13B	76.0	02.1	56.8	06.0	81.6	23.2	82.8	25.5	78.4	22.5	67.0	02.2	76.4	01.7
TowerBase 7B	82.8	25.8	82.1	25.7	83.7	25.9	83.8	25.6	83.0	26.3	82.5	26.4	82.0	26.3
TowerBase 13B	82.7	27.1	83.5	27.3	84.2	27.8	85.0	28.9	83.4	27.2	74.9	23.9	78.2	25.8
ALMA 7B	82.9	24.8	77.1	15.7	82.3	23.0	83.4	25.3	82.4	23.4	82.7	22.7	83.7	24.5
ALMA 13B	83.8	26.2	73.7	17.3	83.2	24.9	84.3	27.1	73.7	25.6	83.6	25.6	83.4	27.1
TowerInstruct 7B	84.8	27.3	84.4	26.6	84.8	27.0	85.2	27.5	84.4	26.4	84.7	27.0	85.0	27.1
TowerInstruct 13B	85.1	28.4	84.8	27.2	85.2	28.0	85.6	29.1	84.9	27.5	85.1	27.8	85.4	28.6
EN→FR														
Concat Enc-Dec	77.8	35.8	68.2	28.9	77.3	35.4	77.5	36.0	–	–	–	–	–	–
NLLB 3.3B	84.8	38.5	–	–	–	–	–	–	–	–	–	–	–	–
EuroLLM-9B-Inst	**86.4**	**40.8**	85.9	40.3	**86.5**	**41.3**	**86.8**	**43.4**	**86.2**	40.5	**86.3**	**41.4**	**86.7**	**42.8**
Llama-2 7B	81.6	33.2	29.5	01.2	81.8	29.6	82.6	34.7	80.9	31.6	82.0	31.5	82.5	30.9
Llama-2 13B	77.0	17.1	54.7	04.2	83.8	35.5	84.5	38.4	81.1	34.2	81.9	20.7	83.4	06.3
TowerBase 7B	84.7	39.9	83.8	37.1	79.0	10.8	78.7	36.2	84.4	40.0	79.1	13.6	76.5	35.4
TowerBase 13B	79.4	39.5	84.9	**41.0**	85.1	40.7	85.9	41.9	85.1	**40.7**	85.4	40.6	69.3	31.7
ALMA 7B	80.8	28.7	52.2	07.1	80.4	25.7	81.1	27.9	80.3	28.9	80.5	27.4	81.3	30.5
ALMA 13B	83.0	33.7	60.0	10.0	82.8	32.7	83.4	33.1	82.9	33.9	82.9	33.9	83.7	35.1
TowerInstruct 7B	85.8	38.1	85.5	35.4	83.4	33.0	86.0	39.6	85.4	36.1	84.1	36.9	85.9	39.1
TowerInstruct 13B	86.2	40.0	**86.0**	39.3	86.0	40.3	86.4	40.9	86.0	39.5	86.0	40.2	86.2	40.8

Table 2: Translation performance (COMET and BLEU) on IWSLT2017, with random, structurally perturbed and gold context, for the prompts considered. **The best value** per column is marked in Bold blue numbers while red marks the second best value; (**) marks best overall. Enc-Dec is short for the encoder-decoder transformer model.

	sentence			random			perturbed			gold		
	COMET	GPRO	CPRO	COMET	GPRO	CPRO	COMET	GPRO	CPRO	COMET	GPRO	CPRO
EN→DE												
Concat Enc-Dec	66.2	**41.7**	46.4	**61.5**	**32.6**	45.3	**66.9**	**53.5**	**60.4**	**67.0**	**56.2**	**60.4**
NLLB 3.3B	**72.3**	41.6	32.0	–	–	–	–	–	–	–	–	–
EuroLLM-9B-Inst	61.5	29.7	**54.7**	50.9	24.5	51.0	41.6	21.8	47.7	43.7	29.6	51.4
Llama-2 7B	35.0	09.7	45.2	27.6	02.3	46.3	39.3	22.1	46.9	41.6	26.1	49.9
Llama-2 13B	34.2	07.6	45.1	28.0	03.0	45.9	40.1	25.5	49.6	42.7	31.1	56.7
TowerBase 7B	39.6	14.1	46.7	35.0	11.2	45.7	44.0	25.1	47.9	45.9	28.9	50.8
TowerBase 13B	56.6	30.8	46.6	31.8	06.6	46.4	51.6	27.3	49.9	50.2	32.2	53.8
ALMA 7B	52.4	22.1	46.4	30.7	06.8	45.8	46.5	25.6	47.2	49.0	30.6	49.9
ALMA 13B	55.3	24.6	46.9	30.3	05.7	47.5	46.3	29.7	52.2	48.6	35.5	58.5
TowerInstruct 7B	57.0	29.9	49.8	40.7	14.5	58.0	53.9	27.1	48.5	55.2	30.7	51.9
TowerInstruct 13B	56.6	30.8	54.5	53.8	21.8	**59.2**	51.6	27.8	55.0	60.9	32.2	59.9
EN→FR												
Concat Enc-Dec	66.5	51.7	76.5	**62.7**	**51.6**	**76.2**	**66.8**	**57.7**	80.5	**67.0**	**65.0**	86.0
NLLB 3.3B	**76.3**	**64.0**	36.9	–	–	–	–	–	–	–	–	–
EuroLLM-9B-Inst	58.5	34.2	06.7	28.8	00.7	17.0	43.2	25.4	11.6	46.9	36.7	13.2
Llama-2 7B	38.0	12.9	90.0	28.7	01.5	64.6	41.9	24.8	64.5	46.1	34.0	68.2
Llama-2 13B	34.1	6.3	89.4	29.1	02.2	49.0	42.5	25.6	59.2	47.1	35.1	63.6
TowerBase 7B	41.5	14.7	**94.5**	38.5	09.8	70.2	45.7	26.7	**85.9**	50.2	36.3	**88.1**
TowerBase 13B	38.0	10.1	78.3	33.7	05.7	74.3	47.6	28.4	80.1	52.5	38.3	82.1
ALMA 7B	42.6	14.7	11.2	29.1	02.4	05.4	41.7	22.7	09.0	45.4	29.7	10.6
ALMA 13B	45.0	16.5	09.4	30.1	03.0	05.3	44.4	26.7	08.3	48.6	34.4	09.8
TowerInstruct 7B	56.6	35.9	55.1	34.9	04.0	23.8	50.3	29.3	52.6	55.1	39.5	56.5
TowerInstruct 13B	57.0	35.1	11.1	47.9	14.1	04.7	53.1	30.3	12.4	58.1	40.4	13.8

Table 3: This table presents the translation performance measured using COMET, the generative (GPRO) and the contrastive (CPRO) pronoun-resolution accuracies on ContraPro dataset, with random, structurally perturbed and gold context, and generic prompt. Random guessing accuracy: 33.3% EN→DE, 50% EN→FR. **The best value** per column is marked in Bold blue numbers while red marks the second best value; (**) marks best overall. Enc-Dec is short for the encoder-decoder transformer model.

```
Translate the following <src_lang> source text to <tgt_lang>:                    (a)
<src_lang>: <src_sentence> <tgt_lang>:
```

```
<src_lang>: <src context 1> <tgt_lang>: <tgt context 1>                          (b)
<src_lang>: <src context 2> <tgt_lang>: <tgt context 2>
<src_lang>: <src sentence> <tgt_lang>:
```

```
<src_lang>: <src context 1> <tgt_lang>: <tgt context 1>                          (c)
<src_lang>: <src context 2> <tgt_lang>: <tgt context 2>
Given the provided parallel sentence pairs, translate the following
↪  <src_lang> sentence to <tgt_lang>:
<src_lang>: <src sentence> <tgt_lang>:
```

Figure 3: a) sentence-level, b) generic, and c) explicit prompt formats. `tgt context` refers to gold translations.

context setup, the perturbation analysis (performance under the perturbed and random context setups), and the attribution analysis looking at the models' internals.

4.1 Performance With the Gold Context

Overall translation performance. Table 2 shows the translation performance (BLEU, COMET) on IWSLT2017 in the sentence-level baseline setup, the generic prompt setup, and the explicit prompt setups. CHRF results are in a separate table (Table 4) for better readability. We analyze the results of different model categories and summarize the observations and their intuitions in the following paragraphs.

We notice that document-level *generic* prompting improves translation performance of all models over the sentence-level baseline. This is expected since document-level prompting gives the model access to inter-sentential context. Moreover, *explicit* prompting improves instruction-finetuned models' performance, while strong base-models (such as TowerBase 13B) degrade in performance. This is also aligned with expectations of the sensitivity of models to the prompt format (Wu et al., 2024), and it highlights the importance of aligning training and inference prompts. However, as the gains with explicit prompting are not substantial even for instruction-tuned models, we opt for the generic prompt format for the pronoun resolution experiments.

For models under consideration in this work, decoder-only LLMs outperform encoder-decoder models at overall translation. This aligns with previous research findings of the potential of LLMs as a new paradigm for document-level translation (Wang et al., 2023). Interestingly, for both language pairs, EuroLLM-9B-Inst outperforms all models in both prompting formats. In the explicit prompting format, TowerInstruct 13B achieves the second-highest performance, while in the generic format, TowerBase 13B comes in second (for EN→FR). EuroLLM-9B-Inst's recipe of multilingual pretraining and instruction tuning seems to have better effects on improving the translation performance compared to the continued multilingual pretraining and translation-specific fine-tuning of Tower models. ALMA models lag behind Tower models despite both employing a two-step fine-tuning strategy on multilingual and parallel data. This raises the need for a deeper investigation into how various design choices (such as the selection and number of finetuning languages, the choice of datasets, and the configuration of hyper-parameters) influence downstream performance.

Further analyzing Table 2, we observe that Llama-2 13B model has a noticeably low performance with explicit gold context for both language pairs. While surprising at first sight, we argue that as the model is pretrained mainly on English data, it might not be sufficient for this task. We look at the translations produced by the model and find that they are mostly repeated words or outputs in the source language instead of the target language.

Pronoun resolution performance. Table 3 shows the generative and contrastive pronoun accuracy and translation performance (COMET) on ContraPro dataset.

Similar to the overall translation performance, We notice that document-level prompting outperforms sentence-level prompting in pronoun resolution performance. A key finding from this analysis is the contrasting ranking compared to the overall translation performance: both encoder-decoder baselines outperform all LLMs in terms of GPRO and COMET scores. Even with gold context, LLMs' performance remains notably poor, with accuracy at or below the random guessing accuracy (33.3% for EN→DE, and 50% for EN→FR). This suggests that there is room to improve LLMs' translation finetuning to better handle context-dependent discourse phenomena.

However, it is important to note that except for the encoder-decoder transformer model that we trained from scratch, we don't have access to other models' training data, therefore, we cannot guarantee that ContraPro is unseen and thus that the evaluation is fair. In particular, NLLB's performance far above chance at the sentence level may be due to such contamination, as sentence-level evaluation forces it to *guess* the pronoun gender

	Sent.	Genric			Explicit		
	base.	rand.	pert.	gold	rand.	pert.	gold
EN→DE							
Concat Enc-Dec	53.0	50.7	53.0	53.1	–	–	–
NLLB 3.3B	**59.7**	–	–	–	–	–	–
EuroLLM-9B-Inst	59.4	**58.8**	**59.1**	**60.4**	**59.2**	**59.5**	****60.7**
Llama-2 7B	51.2	12.1	51.3	52.2	51.0	52.0	53.3
Llama-2 13B	35.1	17.9	53.5	54.8	52.2	32.5	33.5
TowerBase 7B	56.9	56.7	57.0	56.4	57.1	56.8	56.5
TowerBase 13B	57.8	57.9	58.3	59.1	57.9	51.7	54.8
ALMA 7B	54.8	46.6	53.0	54.8	54.5	54.2	55.4
ALMA 13B	56.6	43.5	55.2	56.8	56.2	56.2	57.4
TowerInstruct 7B	57.9	57.4	57.7	58.1	57.4	57.7	57.9
TowerInstruct 13B	58.9	58.2	58.6	59.4	58.2	58.5	59.1
EN→FR							
Concat Enc-Dec	60.9	56.4	60.9	61.3	–	–	–
NLLB 3.3B	**65.9**	–	–	–	–	–	–
EuroLLM-9B-Inst	65.6	65.2	**66.0**	****67.4**	65.8	**66.4**	**67.3**
Llama-2 7B	59.1	06.5	58.3	60.0	59.0	59.2	59.4
Llama-2 13B	55.6	15.1	61.8	63.2	60.0	59.2	51.7
TowerBase 7B	65.5	64.6	44.2	58.9	65.5	48.4	58.5
TowerBase 13B	64.4	**66.2**	65.9	66.6	**66.0**	65.8	55.2
ALMA 7B	56.6	20.4	54.9	55.8	56.6	55.6	57.9
ALMA 13B	59.9	25.3	59.8	60.4	59.7	60.5	61.4
TowerInstruct 7B	64.2	63.0	62.8	65.2	63.3	64.3	64.9
TowerInstruct 13B	65.2	64.9	65.4	65.9	64.9	65.5	65.6

Table 4: CHRF scores on IWSLT2017 test data for the sentence-level baseline and the random, structurally perturbed and gold context, for the prompts considered. **The best value** per column is marked in Bold blue numbers while red marks the second best value; (**) marks best overall. Enc-Dec is short for the encoder-decoder transformer model.

without antecedent information.

Contrastive evaluation measures the classification accuracy of models which does not necessarily correlate with the generative training objective. As suggested by Post and Junczys-Dowmunt (2023), generative scores are better at discriminating document-level systems compared to contrastive scores, which is what we notice in CPRO results where we see surprising trends, with TowerBase 7B leading in EN→FR and TowerInstruct 13B performing comparably to the Concat Enc-Dec model in EN→DE which doesn't align with their GPRO and COMET performance on the data.

4.2 Perturbation Analysis

Structurally perturbed context. From Table 2, we see that structurally perturbing the context has a minimal impact on overall translation performance. All models exhibit only a slight degradation in BLEU, COMET, and CHRF scores when provided with a perturbed context. However, a closer look at the impact of context perturbation on pronoun resolution performance (Table 3) reveals more pronounced effects. Specifically, there is a notable decrease in GPRO performance, ranging from -5 to -10 points, under perturbed context conditions. Nevertheless, the similar level of performance reduction across models suggests that no model stands out in its ability to leverage context effectively. This can be attributed to the fact that none of the models are explicitly trained for context utilization.

Random context. Looking at models' performance with total random tokens, we find that on IWSLT data, EuroLLM-9B-Inst and Tower models (the best at translation) are robust to random context and only degrade slightly in performance, aligning with previous observations of the minimal effect of context perturbation on translation performance. Additionally, those models (except EuroLLM-9B-Inst for EN→FR) show the least difference in GPRO performance between gold and random context setups among all LLMs. Robustness to total random context can be linked to lack of proper context utilization. Although the TowerBlocks dataset used to finetune TowerInstruct models includes context-aware data (as per the dataset card[2]), we hypothesize that general fine-tuning alone may not be sufficient for improving discourse phenomena performance. Explicit, context-aware fine-tuning might be required to ad-

[2]https://huggingface.co/datasets/Unbabel/TowerBlocks-v0.2

dress these challenges effectively.

Further analyzing Table 2, it's noteworthy that the TowerBase 7B model performs better with random context as compared to gold context, even though the latter resembles a few-shot learning scenario (Reinauer et al., 2023). That said, we point out that its translation performance is suboptimal, as it is an intermediate model between the base model Llama-2 7B and the instruction-tuned model TowerInstruct 7B designed specifically for translation.

4.3 Attribution Analysis

We analyze models' internals to see how much the relevant context contributes to the outputs. Figures 4a and 4b present attribution percentages of antecedent tokens (relevant context) as well as of the whole context using ALTI-Logit and input-erasure methods, respectively.

Looking at both attribution methods, we see that for EuroLLM-9B-Inst and TowerInstruct 7B (the best two models at translation among the 5 models tested) antecedent tokens have the lowest attribution percentage to the output. Even though for the TowerInstruct 7B model, overall context tokens have the highest attribution percentage. This suggests that there is a need to explicitly finetune translation LLMs to focus on *relevant context* at inference time.

However, unlike the larger differences in *relevant context* and overall context attributions observed for encoder-decoder models by Mohammed and Niculae (2024), we find no striking differences or clear patterns between the contributions for LLMs. This might be due to the fact that the models have similar backbone structures.

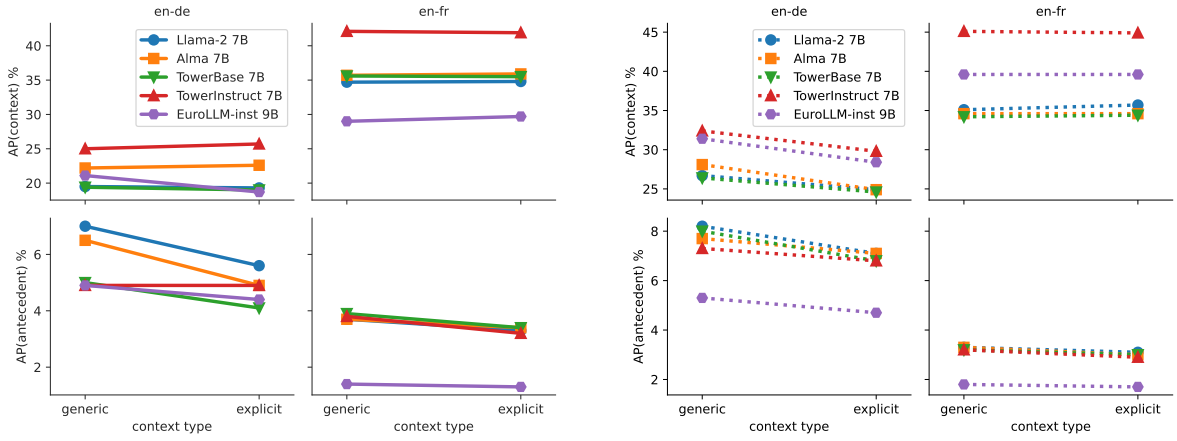
5 Related Work

Context utilization assessment. Works on assessing context utilization in machine translation include the work of Sarti et al. (2023), who build an end-to-end interpretability framework to quantify the plausibility of context-aware encoder-decoder machine translation models. They leverage model internals to contrastively identify context-sensitive target tokens in generated texts and link them to contextual cues justifying their prediction. Using their approach, they were able to consistently detect context-sensitive tokens and their disambiguating rationales across several datasets, models and discourse phenomena. Inspired by this line of research, we evaluate context utilization of LLMs as a possible new paradigm for context-aware translation.

Perturbation and attribution analysis. There are several works that used attribution and perturbation techniques to understand the inner workings of translation LLMs, mostly focusing on the in-context learning (ICL) paradigm —a setup where LLMs "learn" to perform new tasks during inference by being provided with few task demonstrations in the input prompt. Zaranis et al. (2024) use input attribution methods (ALTI) to examine context contributions in translation LLMs within the ICL paradigm. Their findings indicate that the source segments of few-shot examples contribute more significantly than their corresponding target segments, parallel-data fine-tuning alters contribution patterns, and context contributions exhibit a positional bias. Raunak et al. (2023) perturb in-domain translations to better understand their role in ICL. They perform asymmetric perturbation of source-target mappings and find that target perturbations has more negative effect on the translation performance compared to source perturbations. Zhu et al. (2024) also perturb the in-context examples by providing unrelated task (summarization) examples and find that LLMs are not sensitive to the perturbation. Our work combines both interpretability techniques (perturbation and attribution methods) and focuses on context-aware translation task.

LLMs for document-level machine translation. The line of research on adapting LLMs for document-level translation using techniques like LLMs fusion with translation models (Petrick et al., 2023), finetuning LLMs on parallel document-level data (Wu et al., 2024), or a mix of sentence-level and document-level data (Li et al., 2024), generally evaluates on translation metrics and discourse phenomenon accuracy. We complement such evaluations with a finer grained strategy that focuses on the role of context.

Gender bias. Although gender bias does not directly impact our analysis of pronoun resolution —, given that the referents in the ContraPro data are common nouns with clear grammatical gender and, in most cases (the entire German dataset and at least half of the French dataset), are non-human—we recognize that gender bias remains a significant



(a) ALTI-Logit **(b)** input-erasure

Figure 4: Attribution percentages assigned to antecedent tokens (relevant context) and the entire context tokens when force-decoding the correct pronoun in ContraPro data. (a) shows results from ALTI-Logit and (b) displays results from input-erasure attribution methods.

concern for machine translation models and LLMs, as widely explored in research (Rudinger et al., 2018; Zhao et al., 2018; Currey et al., 2022; Rarrick et al., 2023)

6 Conclusion

We use interpretability tools (perturbation and attribution techniques) to analyze LLMs' context-utilization in document-level translation. Our experiments suggest that multilingual pretraining and translation-specific finetuning of LLMs pushes state-of-the-art translation performance beyond encoder-decoder models. However, we highlight that looking at discourse phenomena performance, LLMs show room for improvement. We argue that more care is needed before adopting LLMs as the new standard for document-level translation, and more focused evaluation must be considered. Future research directions include enhancing context-aware translation capabilities of LLMs, potentially through explicit finetuning, and creating datasets with supporting-context annotations for other discourse phenomena to enable extending context-utilization analysis to those phenomena.

7 Limitations

Even though some API-only LLMs (GPT-3.5 and GPT-4) show significant translation improvement compared to encoder-decoder document-level transformers and commercial translation systems (Wang et al., 2023), our analysis approach relies on access to model internals in order to be able to compute attributions of input tokens. Thus, we only used open-source LLMs in our study.

Based on the availability of datasets with context-dependent linguistic phenomena that include supporting context annotations (ContraPro), we experimented only on EN→DE and EN→FR. These two languages belong to the same language family; we leave it to future work to experiment on general translation on other language families.

We chose well-established evaluation metrics in the literature to assess pronoun resolution accuracy. However, we acknowledge the limitations of those metrics. The contrastive metric (CPRO) is not aligned with the generative training objective of models and the generative metric (GPRO) misses cases where the model generates the correct pronoun in a different location in the sentence than the target location.

Due to computational constraints, we were only able to perform the attribution analysis on a small set of models. We hope our work inspires more research into understanding the inner-workings of translation models in context utilization.

For all models except the transformer encoder-decoder model trained from scratch, we do not have details about their training data. This trend of releasing and building on models with secret training data is concerning because it makes fair evaluation impossible.

In our work, we focused on a fine-grained evaluation of context use on a specific phenomenon. Nonetheless, pretrained context-aware metrics could offer more accurate insights into overall models' performance on context use.

8 Ethics Statement

Nowadays, machine translation is a widely adopted technology, sometimes in sensitive, high-risk settings. Even though we propose a fine-grained approach to assessing context utilization, and highlight its importance as we see that improvements in translation performance does not necessarily reflect in discourse phenomena performance, we still rely on automatic evaluation which is imperfect. For systems deployed in critical scenarios, we believe a nuanced case-by-case evaluation is always necessary.

Acknowledgements

We would like to thank Evgenia Ilia, Bryan Eikema and Di Wu for their valuable comments and discussions about this work.

This work is part of the UTTER project, supported by the European Union's Horizon Europe research and innovation programme via grant agreement 101070631. VN also acknowledges support from the Dutch Research Council (NWO) via VI.Veni.212.228.

References

Duarte M. Alves, José Pombal, Nuno Miguel Guerreiro, Pedro Henrique Martins, João Alves, M. 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. *CoRR*, abs/2402.17733.

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, EAMT 2012, Trento, Italy, May 28-30, 2012*, pages 261–268. European Association for Machine Translation.

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, 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. *CoRR*, abs/2207.04672.

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.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. 2022. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8698–8714. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. 2023. Explaining how transformers use context to build predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5486–5513. Association for Computational Linguistics.

Dan Jurafsky and James H. Martin. 2009. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd Edition*. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. 2022. The disagreement problem in explainable machine learning: A practitioner's perspective. *CoRR*, abs/2202.01602.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: quantifying the carbon footprint of computation. *Advanced science*, 8(12):2100707.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.

Yachao Li, Junhui Li, Jing Jiang, and Min Zhang. 2024. Enhancing document-level translation of large language model via translation mixed-instructions. *CoRR*, abs/2401.08088.

António V. Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020*, pages 225–234. European Association for Machine Translation.

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.

Wafaa Mohammed and Vlad Niculae. 2024. On measuring context utilization in document-level MT systems. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 1633–1643. Association for Computational 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, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 61–72. Association for Computational Linguistics.

Bonnie Nash-Webber, Andrei Popescu-Belis, Katja Markert, and Jörg Tiedemann, editors. 2013. *Proceedings of the Workshop on Discourse in Machine Translation, DiscoMT@ACL 2013, Sofia, Bulgaria, August 9, 2013*. 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: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. 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. ACL.

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, WMT 2023, Singapore, December 6-7, 2023*, pages 375–391. Association for Computational Linguistics.

Andrei Popescu-Belis, Sharid Loáiciga, Christian Hardmeier, and Deyi Xiong, editors. 2019. *Proceedings of the Fourth Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2019, Hong Kong, China, November 3, 2019*. 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.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation. *CoRR*, abs/2304.12959.

Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. Gate: A challenge set for gender-ambiguous translation examples. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 845–854.

Vikas Raunak, Hany Hassan Awadalla, and Arul Menezes. 2023. Dissecting in-context learning of translations in gpts. *CoRR*, abs/2310.15987.

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.

Raphael Reinauer, Patrick Simianer, Kaden Uhlig, Johannes EM Mosig, and Joern Wuebker. 2023. Neural machine translation models can learn to be few-shot learners. *arXiv preprint arXiv:2309.08590*.

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.

Gabriele Sarti, Grzegorz Chrupala, Malvina Nissim, and Arianna Bisazza. 2023. Quantifying the plausibility of context reliance in neural machine translation. *CoRR*, abs/2310.01188.

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, Dublin, Ireland, May 22-27, 2022*, pages 3537–3548. 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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura,

Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16646–16661. Association for Computational Linguistics.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George F. Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *CoRR*, abs/2401.06468.

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. *CoRR*, abs/2309.11674.

Emmanouil Zaranis, Nuno Guerreiro, and André F. T. Martins. 2024. Analyzing context contributions in llm-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 14899–14924. Association for Computational Linguistics.

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.

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, Mexico City, Mexico, June 16-21, 2024*, pages 2765–2781. Association for Computational Linguistics.

A Sustainability statement

Our experiments with 13B parameter models run in 95h on 2 GPUs NVIDIA A100 PCIe, and draw 81.69 kWh. Based in the Netherlands, this has a carbon footprint of 30.58 kg CO2e, which is equivalent to 2.78 tree-years. For all other models, the experiments run in 502h on 1 GPU NVIDIA A100 PCIe, and draw 222.08 kWh. Based in the Netherlands, this has a carbon footprint of 83.13 kg CO2e, which is equivalent to 7.56 tree-years (Lannelongue et al., 2021).

Context-Aware Monolingual Human Evaluation of Machine Translation

Silvio Picinini
eBay
spicinini@ebay.com

Sheila Castilho
Salis/Adapt Centre
Dublin City University
sheila.castilho@dcu.ie

Abstract

This paper explores the potential of context-aware monolingual human evaluation for assessing machine translation (MT) when no source is given for reference. To this end, we compare monolingual with bilingual evaluations (with source text), under two scenarios: the evaluation of a single MT system, and the comparative evaluation of pairwise MT systems. Four professional translators performed both monolingual and bilingual evaluations by assigning ratings and annotating errors, and providing feedback on their experience. Our findings suggest that context-aware monolingual human evaluation achieves comparable outcomes to human bilingual evaluations, and suggest the feasibility and potential of monolingual evaluation as an efficient approach to assessing MT.

1 Introduction

MT evaluation has traditionally depended on comparing the translated text with its corresponding source text to assess the MT performance. However, several works in the area have explored the potential of monolingual evaluation (Callison-Burch, 2005; Koehn, 2010; Mitchell et al., 2013; Schwartz, 2014; Fomicheva and Specia, 2016; Graham et al., 2017).

Building on earlier studies suggesting that monolingual post-editors, even without access to the source text, can enhance MT output quality (Callison-Burch, 2005; Koehn, 2010; Mitchell et al., 2013), Schwartz (2014) examined the effectiveness of monolingual post-editing performed by domain experts in improving MT quality and reducing reliance on bilingual post-editing. The author reported that a subject-matter expert monolingual post-editor confidently corrected 87% of sentences based solely on the target text, with 96% of these corrections deemed appropriate following bilingual verification. While our work presented here differs in focus, exploring monolingual evaluation rather than post-editing, it highlights a complementary strength: the importance of context-awareness.

The work of Graham et al. (2017) investigated various evaluation approaches with monolingual evaluation, such as the inclusion or exclusion of reference translations; the impact of additional contextual information (e.g., surrounding sentences); and the influence of displaying metadata to annotators. They concluded that monolingual evaluations, even when crowd-sourced, could effectively measure MT system performance.

Recent advancements in context-aware evaluation methodologies have introduced a deeper focus on the appropriateness and coherence of translations within broader textual contexts (Castilho, 2020, 2021; Castilho et al., 2020; Freitag et al., 2021). These approaches challenge traditional evaluation methodologies that rely on isolated sentences, highlighting the limitations of sentence-level assessments. By integrating contextual information, evaluators can better capture nuanced aspects of translation quality, such as consistency, discourse coherence, and appropriateness within the larger narrative.

Building on this, our study investigates the potential of monolingual assessments in the context, addressing the following research question: *Are monolingual assessments of MT comparable to bilingual assessments?* To explore this, we examine two key scenarios:

a) Can context-aware monolingual assessments of a single MT output provide results comparable to bilingual assessments?

b) Can context-aware monolingual assessments of two different MT outputs (pairwise comparison) achieve comparability with bilingual assessments?

The research question with the two scenarios are

Proceedings of Machine Translation Summit XX Volume 1, pages 138–149, June 23-27, 2025

	Single MT	Pairwise MT	Test Set
DELA	23 (303 w)	30 (576 w)	1, 3
Customer Support	18 (429 w)	53 (750 w)	2, 4

Table 1: Number of sentences and word count from each source per task, along with their associated Test Sets.

illustrated in figure 1. A positive answer to the question would open potential benefits for facilitating the evaluation process.

2 Methodology

2.1 Test Sets

This evaluation was performed using a sentence-in-context format, where the evaluation is at the sentence level, but the evaluator has access to the full context of the document. In one content (travel review), full content is one entire review by a traveler. In the other content (customer support help pages), full content is one entire page explaining a topic to a customer (for example, "listing policies" or "how to buy as a guest"). For this, we used two documents from the Review section of the DELA corpus (Castilho et al., 2021) and two documents from Anonymized's Customer Support pages.[1] The DELA Corpus is a document-Level corpus, in Brazilian Portuguese (the target language that we use in this work) annotated with context-related issues such as ellipsis, gender, lexical ambiguity, number, reference, and terminology. The corpus covers six domains and we chose Reviews (in our case, travel reviews). We constructed 4 different Test Sets. For the scenario where a single MT system output is assessed (Single MT) we used 41 sentences (732 words), 23 from DELA (Test Set 1) and 18 from eBay's Customer Support pages (Test Set 2). For the scenario where two MT systems are assessed (pairwise MT), we used 83 sentences (1326 words), 30 from DELA (Test Set 3) and 53 from eBay (Test Set 4). Number of sentences and word counts are in table 1.

2.2 Systems, Evaluators and Language

The MT systems used for the evaluation were the freely available Microsoft (MT1) and DeepL (MT2). We note that only MT1 (Microsoft) is used in the Single MT scenario.[2]

The language pair of this experiment was English (EN) as the source and Brazilian-Portuguese (pt-BR) as the target. Four in-house professional translators participated in the evaluation. They are native speakers of pt-BR and long-time translators for eBay and for a major language services provider.

2.3 Metrics

Evaluators assessed the systems' output in terms of:

- Overall Ratings (Likert scale from 1-5)
- MQM Error Annotation

Ratings: these Ratings, used at eBay in-house for MT performance evaluation, is measured in a 1-5 scale and considers both adequacy and fluency. It is used to evaluate the experience and success of an end-user in understanding the final MT output, and also the experience of a translator who receives the MT output as an input for post-editing. The reason for using this in-house Ratings is to maintain ecological validity, as it reflects the Ratings these translators are familiar with. Additionally, since eBay will continue using these Ratings in future evaluations, the results here can provide valuable insights for the company's ongoing comparisons.

The scale consists of:

1. Critical/Incomprehensible: When the translation:
    - is incomprehensible (hallucinations, meanings that make no sense in the text)
    - is completely not in the target language
    - contains critical errors
    - can be a misleading mistranslation
    - may carry health, safety, legal, reputation, religious or financial implications
    - contains profanities/insulting words not appropriate in the context
2. Significant errors/effort: Translation contains significant errors and may be incomplete (such as significant mistranslations, significant portions untranslated, significant omissions) and requires significant effort to understand it. It would require a significant amount of change to be usable.

[1] Document 1 is "Buying as a Guest", available at url anonymized, and document 2 is "Listing Policies", available at url anonymized

[2] MT translations were obtained on 22 February 2024 from the public versions of the MT providers on their sites.

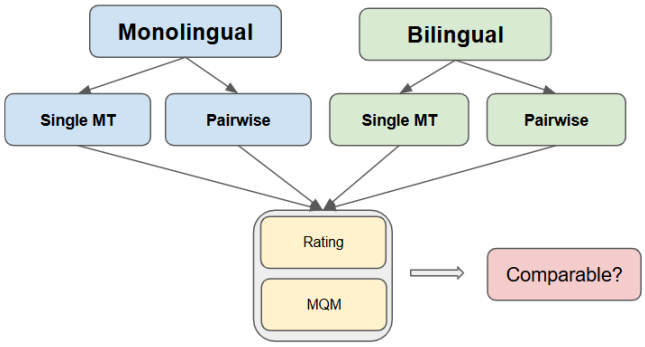


Figure 1: Design of the tasks to answer the RQ. "Single MT" refers to the scenario where only one MT output is assessed, whereas "Pairwise" refers to the scenario where two MT systems are assessed. Both scenarios are assessed in the two modalities, monolingual and bilingual, in terms of Ratings and Error Annotation.

3. Some errors/some effort: There are errors in translation, such as untranslated words and mistranslations (wrong but not misleading translations). The translation requires some effort to understand it and change it.

4. Do not affect understanding: There are minor grammar errors: wrong agreement, word order, missing prepositions, minor omissions or other errors that do not affect understanding. It would require only minor changes.

5. Complete/fluent: Translation is (or seems to be) correct, complete, fluent, easy to read, and contains no errors. It looks like (almost) human translation and requires no changes.

Error Annotation: Evaluators annotated errors using the MQM error typology (Lommel et al., 2014), following the methodology and guidelines outlined in (Freitag et al., 2021) (see Appendices A and B).

2.4 Setup

The evaluation was conducted using an online spreadsheet sent to the evaluators. As mentioned earlier, the assessment was performed in context, with access to the full text in two modalities:
a) ***Monolingual***, where only the MT output was provided - Task 1.
b) ***Bilingual***, where both the MT output and source text were provided - Task 2.

Evaluators performed both Ratings evaluations and Error Annotations (see Section 2.3) across these two modalities (Monolingual and Bilingual), in two scenarios: Single MT and Pairwise MT. This resulted in a total of 16 evaluations.

The evaluation process followed a structured sequence, with Task 1 (Monolingual) always performed first, followed by Task 2 (Bilingual), with a one-week interval between tasks. This order was designed to minimize the likelihood of evaluators recalling source text information from the bilingual evaluation. By conducting the monolingual evaluation first, evaluators identified visible errors in the target text without reference to the source. In the subsequent bilingual evaluation of the same text, evaluators could incorporate errors previously identified during the monolingual assessment and, additionally, detect new errors that were only apparent through comparison with the source text. The goal of this approach was to determine how many errors identified during the bilingual evaluation had already been noted in the monolingual assessment and how many new errors were detected only through bilingual evaluation.

Moreover, we note that the order of Test Sets and scenarios were randomized. The distribution and randomization of the evaluations can be seen in table 2.

3 Results

This section presents the outcomes of the experiments, organized by the tasks performed. The first part, Section 3.1, details the results from Task 1, carried out in the **Monolingual** modality, while Section 3.2 focuses on Task 2, conducted in the **Bilingual** modality. The results for both tasks are reported across the two scenarios - *Single MT* and *Pairwise MT* - covering the overall Ratings and Error Annotation.

Modality	Evaluator 1		Evaluator 2		Evaluator 3		Evaluator 4	
	Scenario	Test Set	Scenario	Test Set	Scenario	Test Set	Scenario	Test Set
Task 1 - Monolingual	Pairwise	TS3	Pairwise	TS4	Single	TS1	Single	TS2
		TS4		TS3		TS2		TS1
	Single	TS1	Single	TS2	Pairwise	TS3	Pairwise	TS4
		TS2		TS1		TS4		TS3
Task 2 - Bilingual	Single	TS2	Single	TS1	Pairwise	TS4	Pairwise	TS3
		TS1		TS2		TS3		TS4
	Pairwise	TS4	Pairwise	TS3	Single	TS2	Single	TS1
		TS3		TS4		TS1		TS2

Table 2: Distribution of evaluation tasks by modality (Monolingual - Task 1, and Bilingual - Task 2), scenarios (Single MT and Pairwise MTs), and Test Sets (TS) assigned to each evaluator.

3.1 Task 1 - Monolingual Evaluation

The monolingual results for the Single MT scenario are shown in table 3, while the results for the Pairwise MT scenario are shown in table 4.

Overall Ratings - The overall Ratings for the *Single MT scenario* show a significant variation among evaluators (table 3). Interestingly, results for the *Pairwise MT scenario* show a more consistent Ratings for the MT2. For MT1, there is a wider range of Ratings, with evaluator 4 penalizing MT1 more severely (table 4).

Error Annotation - The results for the Error Annotation are reported according to i) the overall number of errors, and ii) the severity Major. We provide the number of Minor errors only to facilitate the understanding that All Errors is the sum of Major and Minor errors. Looking at i) the overall number of errors, we are assessing the overall ability of the monolingual evaluation of capturing as many of the total errors as the bilingual evaluation. Looking only at Major errors, we incorporate the severity element into the analysis and assess the ability of monolingual evaluation capturing the important errors, in the same way that a bilingual evaluation would.

i) All errors - *Single MT scenario*: We note a wide variation in the reported numbers from evaluators 3 and 4, with evaluators 1 and 2 more in agreement (table 3).
Pairwise MT scenario: Also shows a wide variation in the number of total errors for MT2 and for MT1 (table 4).

ii) Major errors - *Single MT scenario*: When only Major errors are considered, a wide variation can be seen from the same evaluators in the Single MT scenario.
Pairwise MT scenario: Similarly, when only Major errors are considered when evaluating two MTs monolingually, we see for MT2 a fair variation, whereas for MT1 a wider variation is observed.

3.2 Task 2 - Bilingual Evaluation

This subsection presents the results for the Bilingual modality, covering both the overall Ratings evaluation and Error Annotation. The bilingual results for the Single MT scenario are shown in table 5, while results for the Pairwise scenario are shown in table 6.

Overall Ratings - We note that Ratings for the *Single MT scenario* in the bilingual modality show a fair variation among evaluators (table 5). *The Pairwise MT scenario* results show more consistent Ratings for the MT2. For MT1, there is a wider range of Ratings (table 6).

Error Annotation - i) All errors - *Single MT scenario*: We note a wide variation in the reported numbers from evaluators 3 and 4, with evaluators 1 and 2 more in agreement (table 5).
Pairwise MT scenario: Also shows a wide variation in the number of total errors for MT2 and for MT1 (table 6).

ii) Major errors - *Single MT scenario*: When only Major errors are considered, a wide variation can be seen from the same evaluators in the Single MT scenario.
Pairwise MT scenario: Similarly, when only Major errors are considered when evaluating two MTs bilingually, we see for MT2 a fair variation, whereas for MT1 a wider variation is observed.

4 Analysis of the Results

The aim of this evaluation is to compare the extent to which MT systems can be assessed monolin-

Task 1 Single MT	Ratings	Error Annotation		
		All errors	Major	Minor
Eval 1	3.8	23	8	15
Eval 2	3.7	20	9	11
Eval 3	4.4	6	2	4
Eval 4	2.9	46	26	20
Average	3.7	24	11	13

Table 3: Task 1 - Monolingual results for Overall Ratings and Error Annotation for the **Single MT** scenario.

Task 1 Pairwise MT	Ratings		Error Anotation					
	MT1	MT2	MT1			MT2		
			All errors	Major	Minor	All errors	Major	Minor
Eval 1	4.0	4.9	44	28	16	9	3	6
Eval 2	4.2	4.8	29	13	16	10	2	8
Eval 3	4.1	4.7	28	20	8	10	4	6
Eval 4	3.0	4.4	71	45	26	30	10	20
Average	3.8	4.7	43	27	17	15	5	10

Table 4: Task 1 - Monolingual results for Overall Ratings and Error Annotation for the **Pairwise MT** scenario.

Task 2 Single MT	Rating	Error Annotation		
		All errors	Major	Minor
Eval 1	4.2	23	6	17
Eval 2	3.6	26	7	19
Eval 3	4.0	17	9	8
Eval 4	3.3	48	15	33
Average	3.8	29	9	19

Table 5: Task 2 - Bilingual results for Overall Ratings and Error Annotation for the **Single MT** scenario.

Task 2 Pairwise MT	Ratings		Error Annotation					
	MT1	MT2	MT1			MT2		
			All errors	Major	Minor	All errors	Major	Minor
Eval 1	4.1	4.9	41	27	14	6	1	5
Eval 2	4.3	4.7	13	6	7	5	1	4
Eval 3	4.0	4.6	36	24	12	10	3	7
Eval 4	3.1	4.3	81	40	41	32	10	22
Average	3.8	4.6	43	24	19	13	4	10

Table 6: Task 2 - Bilingual results for Overall Ratings and Error Annotation for the **Pairwise MT** scenario.

gually to the same (or some) extent as bilingually. This section analyses whether the results of these experiments answer our RQ in the two scenarios explored: Single MT and Pairwise MT.

4.1 Are monolingual results comparable to bilingual results when evaluating a single MT?

Overall Ratings - Despite the variation in the overall Ratings, monolingual and bilingual evaluation of a single MT system seem to be comparable. The average scores of the four evaluators are close, at 3.7 for the monolingual task and 3.8 for the bilingual task, indicating that the monolingual evaluation may be as effective as the bilingual evaluation when evaluating a single MT.

Error Annotation - *All errors*: The results for the monolingual and bilingual tasks are relatively close when considering all errors tagged. The average number of all errors is 24 in the monolingual task, and 29 errors in the bilingual task, a difference of 17%. *Major errors*: Similarly, the results for the monolingual and bilingual tasks are relatively close when considering only Major errors tagged. The average results are close, with an average of 11 Major errors in the monolingual task, and 9 in the bilingual task.

These results show that both monolingual and bilingual evaluations are capturing enough severe errors to reflect the performance of one MT. This indicates that a monolingual evaluation may be effective when using the MQM typology, where *severity* is taken into account.

4.2 Are monolingual results comparable to bilingual results when evaluating two different MTs?

Overall Ratings - When evaluating two MT systems, both monolingual and bilingual evaluation seem to agree very closely. In both tasks, evaluators agree that MT2 performs well, with close average Ratings of 4.7, in the monolingual task, and 4.6 in the bilingual task. For MT1, even with a wider variation in Ratings, and also a stricter view of performance from evaluator 4, the average Ratings of four evaluators is very close, with 3.8 for both monolingual and bilingual. Both the monolingual and the bilingual evaluations established that MT2 outperforms MT1, indicating the usefulness of the monolingual evaluation.

Error Annotation - *All errors*: The results for the monolingual and bilingual tasks are close, with MT2 showing 15 errors in the monolingual and 13 in the bilingual tasks; and MT1 showing 43 errors in both monolingual and bilingual tasks. The difference of 2 errors in the MT2 assessment corresponds to 13% only. *Major errors*: Similarly, when only Major errors are considered, the results for the monolingual and bilingual tasks are close. The average number of Major errors tagged in MT2 is 5 for monolingual and 4 for bilingual, while in MT1 is 27 for monolingual and 24 for bilingual, also relatively close.

Both the monolingual and the bilingual evaluations are able to capture similar number of errors and severity, and establish that MT2 outperforms MT1. Both monolingual and bilingual are capturing enough severe errors to reflect the performance of the translation, indicating that a monolingual evaluation may be as effective as the bilingual evaluation in terms of Error Annotation for two MT systems.

4.3 Can monolingual assessments of MT be effective?

Following the results of the two evaluation tasks, we performed an analysis of the systems' outputs in order to identify what errors could and could not be detected monolingually, and also why they could be detected.

- **Most errors appear to be detectable monolingually (from target only)**

Analyzing all reported errors and categorizing them as either "detectable monolingually" or "not detectable monolingually" revealed that the vast majority (98%) seem to be identifiable solely from the target language (see Table 7). This finding aligns with previous results demonstrating that monolingual evaluation seems to yield comparable insights to bilingual evaluation.

Monolingual	Detected	Not Detected	% Detected
Minor	182	2	98.91%
Major	143	6	95.97%
Total	325	8	97.60%

Table 7: Monolingual detectability results. "Detected" refers to errors identified monolingually, "Not Detected" refers to errors missed monolingually, and "% Detected" is the percentage of errors detected monolingually for each error type.

- **Many errors seem to be detectable from the context of the target only**

The impact of context in identifying errors is noticeable. One text from Test Set 2 (see Table 1) discusses the Great Wall of China. In Brazilian Portuguese (pt-BR), the word "wall" has three distinct translations: **parede** (for interior walls), **muro** (for perimeter walls around properties), and **muralha** (for defensive walls surrounding cities or fortresses, as in the case of the Great Wall). One of the MT systems inconsistently used all three translations, resulting in clear mistranslations that became particularly apparent when viewed in context:

*"Agora sobre o **muro** em si. Mutianyu é a seção mais longa e totalmente restaurada da Grande Muralha aberta aos turistas. Existem 23 torres de vigia, cerca de uma a cada cem metros em uma montanha ascendente e eu quero dizer realmente eles são íngremes no S***! Perdoem a minha linguagem, mas maldita a maioria de nós estava cansada em cobrir apenas 5 desses. Ambos os lados da **muralha** têm um parapeito crenado para que os soldados pudessem disparar flechas contra o inimigo em ambos os lados. Isso é muito raro em outras seções da **parede**".*

Another type of error that became noticeable through context was grammatical gender. For example, in a text about the Notre Dame cathedral (Test Set 1), several mistranslations arose due to gender mismatches, as the word catedral ("cathedral") is feminine in Portuguese. In the example below, the pronoun ele (masculine) should have been ela (feminine) to correctly refer to Notre Dame:

*Para mim, a única atração absoluta de Paris era **Notre Dame** e eu nem tinha percebido. Fiquei impressionado com o detalhe e a sensação que **ele** deu!*

These errors were identifiable in the monolingual task due to the availability of context, showing that a significant number of errors seem to be detectable through monolingual evaluation when contextual information is present.

- **Some errors seem to be detectable only through bilingual evaluations**

Building on the previous findings (see Table 7), the analysis indicates that only a small proportion of errors require access to the source text for detection. While approximately 98% of errors were identified monolingually, a limited number of mistranslations remained undetected without bilingual evaluation. Examples include:

- The phrase "Find guest order details" was mistranslated as "Find guest order," omitting the key term "details".
- The source sentence "the mountain ridge was steep" was translated as "the mountain ridge was incredible", replacing "steep" with an incorrect adjective, which is not visible without the source.
- The sentence "They had some pan cake places too so it does have more than just Chinese menu" was mistranslated as "cake", entirely omitting the compound "pan cake."
- In the example, "You can usually buy on eBay without an account if the item is selling for less than $5,000 and it's offered with Buy It Now. You need an eBay account to bid on an auction item," the second sentence was omitted entirely from the translation.

In contrast, certain errors were evident from the target text alone, without requiring the source text for detection. For example:

- The sentence "how to buy as a guest on eBay" was translated as "Como comprar como hóspede no eBay." Here, the MT incorrectly translated "guest" as "hóspede" which refers to a hotel guest in Portuguese, instead of the correct term "convidado," meaning a guest for an event or website.
- The phrase "Top Takeaway" was left untranslated, which is a clear error visible in the target language as it remains in English rather than the target language.

These findings show that while bilingual evaluation offers additional layers of verification, the majority of errors were identified through careful monolingual evaluation.

Questions	Task 1	Task 2
Ease of finding error	8.25	9.5
Confidence in the Single MT eval	8.0	9.25
Confidence in the Pairwise MT eval	8.0	9.0
Time required	3.75	6.5
Effort required	7.0	6.5
Satisfaction with evaluation	8.5	9.25

Table 8: Survey results for Monolingual (task 1) and Bilingual (task 2) evaluation.

4.4 Survey

In order to grasp the evaluators' view on both tasks, monolingual and bilingual, a survey was designed. After each task, evaluators answered five questions, on a scale from 0-10 (where 0 is a low score, and 10 a high score) relating to the issues shown in table 8.[3] These results reflect the subjective perception of a limited number of evaluators and are intended to provide some basic information on the experience of the evaluators when doing evaluations in monolingual and bilingual scenarios.

The results suggest some differences between monolingual (Task 1) and bilingual (Task 2) evaluation approaches. Overall, evaluators found the bilingual evaluation easier, with higher scores for the ease of finding errors (9.5 vs 8.25) and greater satisfaction (9.25 vs 8.5). Confidence levels were slightly higher in the bilingual evaluation for single MT assessment (9.25 vs 8.0) and for comparing two MT outputs (9.0 vs 8.0).

Regarding the evaluators' note on time and effort, we note that bilingual evaluations may have required more time (6.5 vs 3.75) due to the need to read and compare the source text alongside the translations. Interestingly, despite taking more time, bilingual evaluations were perceived as requiring less effort (6.5 vs 7.0). This suggests that "less effort" in the bilingual modality may reflect reduced cognitive strain rather than overall time efficiency. Finally, while most results showed slight differences between tasks, a noticeable increase in

[3] See Appendix C for full questions.

time required for bilingual evaluations suggests a trade-off between effort and time efficiency. This finding suggests that less effort does not necessarily equate to less time spent on the task.

5 Conclusion and Future work

This study, while limited in scope, highlights the potential of monolingual evaluation as a practical and effective alternative method for assessing MT performance. The results suggest that monolingual evaluation may be comparable to bilingual evaluation in assessing the performance of one MT output and in comparing multiple MT outputs - an approach commonly used in companies and academia alike.

Monolingual evaluation seems to be particularly effective when using contextual information, which reinforces the importance of assessing translations within context rather than isolated sentences. Monolingual evaluation also seems to show a robust capability for error detection, with approximately 98% of errors identifiable by examining the target language alone. These findings suggest the potential usefulness of monolingual approaches as both a complementary and standalone alternative method for MT evaluation. Additionally, findings from the survey suggest that monolingual evaluation may not significantly differ from bilingual evaluation in terms of translators confidence and satisfaction. Notably, monolingual evaluation may be a faster task. While evaluators reported slightly higher effort — likely due to the need for rereading in the absence of a source text — this increased cognitive effort does not translate to longer task completion times. As a result, monolingual evaluations may offer a more time- and cost-efficient approach to MT assessment without compromising reliability.

By suggesting the potential effectiveness of monolingual evaluation, this paper contributes to the field of MT evaluation and highlights potential benefits. One notable potential advantage is the expansion of the evaluator pool, as monolingual evaluations require proficiency in only one language, making it easier to find qualified evaluators, even for less commonly spoken language pairs. Furthermore, the reliance on monolingual assessment aligns well with the needs of generative AI evaluation, where systems produce text in a single language. This suggests that monolingual evaluation could serve as a valuable framework for assessing text generation tasks beyond MT, including AI-generated content.

Acknowledgments

Our thanks to Katja Zuske for the review and to the evaluators that helped this effort. The second author benefits from being member 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

Chris Callison-Burch. 2005. Linear b system description for the 2005 nist mt evaluation exercise. In *Proceedings of the NIST 2005 Machine Translation Evaluation Workshop*.

Sheila Castilho. 2020. On the same page? comparing IAA in sentence and document level human mt evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1150–1159. Association for Computational Linguistics.

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*, pages 34–45. Association for Computational Linguistics.

Sheila Castilho, João Lucas Cavalheiro Camargo, Miguel Menezes, and Andy Way. 2021. Dela corpus: A document-level corpus annotated with context-related issues. In *Proceedings of the Sixth Conference on Machine Translation*, pages 571–582. Association for Computational Linguistics.

Sheila Castilho, Maja Popović, and Andy Way. 2020. On Context Span Needed for MT Evaluation. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC'20)*, page 3735–3742, Marseille, France.

Marina Fomicheva and Lucia Specia. 2016. Reference bias in monolingual machine translation evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 77–82, Berlin, Germany. 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.

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.

Philipp Koehn. 2010. Enabling monolingual translators: Post-editing vs. options. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 537–545.

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.

Linda Mitchell, Johann Roturier, and Sharon O'Brien. 2013. Community-based post-editing of machine-translated content: monolingual vs. bilingual. In *Proceedings of the 2nd Workshop on Post-editing Technology and Practice*.

Lane Schwartz. 2014. Monolingual post-editing by a domain expert is highly effective for translation triage. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 34–44, Vancouver, Canada. Association for Machine Translation in the Americas.

A Annotation Guidelines

The Annotation Guidelines used are the ones published by (Freitag et al., 2021) and are displayed in table 9.

B Error Categories, Severity and Description

The Error Categories, along with their severity and description, were published by (Freitag et al., 2021) and are displayed in tables 10 and 11.

C Survey questions

The survey questions used can be seen in table 12.

You will be assessing translations at the segment level, where a segment may contain one or more sentences. Each segment is aligned with a corresponding source segment, and both segments are displayed within their respective documents. Annotate segments in natural order, as if you were reading the document. You may return to revise previous segments. Please identify all errors within each translated segment, up to a maximum of five. If there are more than five errors, identify only the five most severe. If it is not possible to reliably identify distinct errors because the translation is too badly garbled or is unrelated to the source, then mark a single Non-translation error that spans the entire segment. To identify an error, highlight the relevant span of text, and select a category/sub-category and severity level from the available options. (The span of text may be in the source segment if the error is a source error or an omission.) When identifying errors, please be as fine-grained as possible. For example, if a sentence contains two words that are each mistranslated, two separate mistranslation errors should be recorded. If a single stretch of text contains multiple errors, you only need to indicate the one that is most severe. If all have the same severity, choose the first matching category listed in the error typology (eg, Accuracy, then Fluency, then Terminology, etc). Please pay particular attention to document context when annotating. If a translation might be questionable on its own but is fine in the context of the document, it should not be considered erroneous; conversely, if a translation might be acceptable in some context, but not within the current document, it should be marked as wrong. There are two special error categories: Source error and Non-translation. Source errors should be annotated separately, highlighting the relevant span in the source segment. They do not count against the five-error limit for target errors, which should be handled in the usual way, whether or not they resulted from a source error. There can be at most one Non-translation error per segment, and it should span the entire segment. No other errors should be identified if Non-Translation is selected.

Table 9: MQM Annotator guidelines

Error Category	**Description**
Accuracy	
-Addition	-Translation includes information not present in the source.
-Omission	-Translation is missing content from the source.
-Mistranslation	-Translation does not accurately represent the source.
-Untranslated text	-Source text has been left untranslated.
Fluency	
-Punctuation	-Incorrect punctuation (for locale or style).
-Spelling	-Incorrect spelling or capitalization.
-Grammar	-Problems with grammar, other than orthography.
-Register	-Wrong grammatical register (eg, inappropriately informal pronouns).
-Inconsistency	-Internal inconsistency (not related to terminology).
-Character encoding	-Characters are garbled due to incorrect encoding.
Terminology	
-Inappropriate for context	-Terminology is non-standard or does not fit context.
-Inconsistent use	-Terminology is used inconsistently.
Style	
-Awkward	-Translation has stylistic problems.
Locale convention	
-Address format	-Wrong format for addresses.
-Currency format	-Wrong format for currency.
-Date format	-Wrong format for dates.
-Name format	-Wrong format for names.
-Telephone format	-Wrong format for telephone numbers.
-Time format	-Wrong format for time expressions.
Other	-Any other issues.
Source error	-An error in the source. Non-translation Impossible to reliably characterize distinct errors.
Non-translation	-Impossible to reliably characterize distinct errors.

Table 10: Error categories with their definitions

Severity	Severity Definition
Major	Actual translation or grammatical errors
Minor	Smaller imperfections
Neutral	Purely subjective opinions about the translation

Table 11: Severities and their definitions

Questions
How easy was it to find errors in the monolingual evaluation? (0 not at all, 10 very easy)
How confident were you in your evaluation when assessing the SINGLE output monolingually? (0 not at all, 10 completely confident)
How confident were you in your evaluation when assessing the TWO MT outputs monolingually? (0 not at all, 10 completely confident)
How much time did the monolingual evaluation require? (0 very little time, 10 a lot of time)
How much effort did the monolingual evaluation require? (0 very little effort, 10 a lot of effort)
How satisfied were you with the monolingual evaluation? (0 not at all, 10 completely satisfied)
How easy was it to find errors in the bilingual evaluation? (0 not at all, 10 very easy)
How confident were you in your evaluation when assessing the SINGLE output bilingually? (0 not at all, 10 completely confident)
How confident were you in your evaluation when assessing the TWO MT outputs bilingually? (0 not at all, 10 completely confident)
How much time did the bilingual evaluation require? (0 very little time, 10 a lot of time)
How much effort did the bilingual evaluation require? (0 very little effort, 10 a lot of effort)
How satisfied were you with the bilingual evaluation? (0 not at all, 10 completely satisfied)

Table 12: Survey questions

Culture-aware machine translation: the case study of low-resource language pair Catalan-Chinese

Xixian Liao, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari
Javier García Gilabert, Miguel Claramunt Argote, Ella Bohman, Maite Melero
Barcelona Super Computing Center (BSC)
{xixian.liao,carlos.escolano,audrey.mash,francesca.delucafornaciari,
javier.garcia1,miguel.claramunt,ella.bohman,maite.melero}@bsc.es

Abstract

High-quality machine translation requires datasets that not only ensure linguistic accuracy but also capture regional and cultural nuances. While many existing benchmarks, such as FLORES-200, rely on English as a pivot language, this approach can overlook the specificity of direct language pairs, particularly for underrepresented combinations like Catalan-Chinese. In this study, we demonstrate that even with a relatively small dataset of approximately 1,000 sentences, we can significantly improve MT localization. To this end, we introduce a dataset specifically designed to enhance Catalan-to-Chinese translation by prioritizing regionally and culturally specific topics. Unlike pivot-based datasets, our data source ensures a more faithful representation of Catalan linguistic and cultural elements, leading to more accurate translations of local terms and expressions. Using this dataset, we demonstrate better performance over the English-pivot FLORES-200 *dev* set and achieve competitive results on the FLORES-200 *devtest* set when evaluated with neural-based metrics. We release this dataset as both a human-preference resource and a benchmark for Catalan-Chinese translation. Additionally, we include Spanish translations for each sentence, facilitating extensions to Spanish-Chinese translation tasks.

1 Introduction

In recent years, the field of neural machine translation (NMT) has seen substantial progress in the development of multilingual models, which can translate across multiple languages as a single unified model (e.g., Zhang et al., 2020; Siddhant et al., 2020; Fan et al., 2021; Costa-jussà et al., 2022; Kudugunta et al., 2024), as well as the creation of human-translated multilingual benchmark datasets (e.g., Costa-jussà et al., 2022; Federmann et al., 2022). These advancements have pushed the boundaries of many-to-many translation capabilities. However, practical applications often require systems to be tailored to specific cultural and regional contexts (e.g., Naveen and Trojovskỳ, 2024). One particularly challenging area is the translation of texts that contain entity names, as cultural-related references can vary significantly across languages (Conia et al., 2024). Translating names between languages with different scripts, such as Latin and logographic (e.g., Chinese), also involves transliteration to maintain ease of pronunciation and closeness to the original sound. Sometimes, the same name can even yield different transliterations based on the source language's pronunciation. For example, the name *José* is transliterated as 若泽(ruò zé) from Portuguese to Chinese, but from Spanish, it becomes 何塞(hé sài). Therefore, we need to adapt the many-to-many system to be more language- and culture-specific.

This study focuses on the Catalan-to-Chinese (CA→ZH) translation, a relatively underexplored area despite its growing relevance given the deepening economic and cultural connections between Catalonia and China. Chinese speakers form one of the five largest immigrant communities in Catalonia, where Catalan is an official language.[1] Besides, China is also Catalonia's third-largest non-European investor and the top source of non-European, non-English-speaking tourists.[2] These growing interactions underline the urgent need for effective translation tools to facilitate communication and foster collaboration between Catalan and Chinese speakers. Despite its significance, developing robust CA-ZH MT systems remains challeng-

[1] https://www.idescat.cat/novetats/?id=4815&lang=en. Accessed January 3, 2025.

[2] https://catalonia.com/w/catalan-government-launches-china-desk-to-promote-chinese-investment-and-strengthen-economic-ties#. Accessed January 3, 2025.

Proceedings of Machine Translation Summit XX Volume 1, pages 150–161, June 23-27, 2025

ing due to the limited availability of high-quality parallel datasets.

In this study, we address the problem of adapting multilingual NMT models to CA→ZH for more region-specific translation. More specifically, the contributions of our work are as follows:

- Human-crafting a Catalan-Chinese parallel dataset containing 1,022 sentences sourced from Catalan/Spanish Wikimedia, translated directly to Mandarin Chinese. This dataset captures cultural and linguistic nuances more specific to Catalonia and Spain than existing benchmark datasets, which are more English-centric.[3]
- Demonstrating the benefits of using preference data with more region-specific content in Contrastive Preference Optimization (CPO) to align the model with human preferences, especially for cultural-specific terms. This approach better enhances the model's ability to handle both region-specific content and English-centric data. Notably, with only 1,022 sentences, we achieve good improvements in MT localization.

2 Related work

2.1 Research on Catalan-Chinese machine translation

Research on CA-ZH MT remains limited, with most previous efforts focusing on creating and mining parallel corpora for this low-resource language pair. Early work by Costa-Jussa et al. (2019) first addressed the lack of resources by creating a pseudo-parallel corpus via pivot translation, with Spanish as the intermediary language. Later, Zhou (2022) created human-selected CA-ZH parallel corpora by mining and validating bitexts from Wikipedia. Their efforts resulted in two datasets: CA-ZH 1.05 (110k sentence pairs) and CA-ZH 1.10 (65k higher-quality pairs). Using these datasets, Liu (2022) fine-tuned the M2M-100-418M multilingual model (Fan et al., 2021). Their full fine-tuning improved translation performance for both CA→ZH and ZH→CA directions, achieving BLEU score gains of +0.3–0.5 with the larger CA-ZH 1.05 corpus and +0.1–0.2 with the smaller, higher-quality CA-ZH 1.10 corpus. More recently, Chen et al. (2024) combined pivot translation (using Spanish) with multilingual training to leverage both synthetic and authentic data. Using the FLORES-200 benchmark (Costa-jussà et al., 2022), their findings showed that fine-tuning M2M-100-418M on the authentic CA-ZH dataset from Zhou (2022) only marginally improved the spBLEU score from 22.0 to 22.4. However, combining pseudo-parallel CA-ZH and Spanish-Chinese (ES-ZH) data alongside authentic CA-ZH and ES-ZH data yielded a significant improvement, increasing the spBLEU score to 26.7.

In this study, we take a different approach by creating a much smaller dataset of authentic CA-ZH data, consisting of 1,022 sentence pairs. Despite the dataset's small size, we demonstrate meaningful improvements in translation performance.

2.2 Contrastive Preference Optimization

Reinforcement Learning from Human Feedback (RLHF) has proven effective in aligning large language models (LLMs) with human preferences (Christiano et al., 2017; Ouyang et al., 2022). However, RLHF relies on a complex training pipeline, requiring first the training of a reward model based on human preference data. To simplify the training, recent work has proposed contrastive preference learning methods, such as Direct Preference Optimization (DPO) (Rafailov et al., 2024), which tune models directly on human preference data without explicitly training a reward model. The primary objective of these methods is to increase the likelihood gap between preferred and dispreferred responses.

Building on DPO, Contrastive Preference Optimization (CPO) was originally developed for machine translation tasks. CPO trains models to consistently favor preferred translations and avoid generating adequate but not perfect outputs. It has demonstrated significant improvements in translation quality. For example, in Spanish-to-Aranese translation tasks using only the FLORES-200 *dev* split, CPO outperformed both supervised fine-tuning and 5-shot fine-tuning, achieving a 1.9 BLEU score improvement with a Qwen2-0.5B-based (Yang et al., 2024) distillation model evaluated on the FLORES-200 *devtest* split (Hu et al., 2024).

In this study, we apply CPO to CA→ZH translation using a preference dataset that captures cultural and linguistic nuances more specific to Catalonia and Spain, in contrast to the more common English-pivot approach (using FLORES-200). We then compare the results to assess the impact of this lo-

[3]The dataset is available upon request from the authors.

calization. In the following section, we describe the construction of the preference datasets.

3 Dataset Construction

CPO requires a preference dataset, consisting of a "prompt", a "chosen" completion, and a "rejected" completion. The objective is to train the model to prefer the "chosen" response over the "rejected" response.

This section describes the construction of our preference datasets. Specifically, we create two datasets to assess the effects of using different types of data in CPO:

- `CPO FLORES DEV`: Based on the *dev* split of the FLORES-200 dataset, which is an English-pivot multilingual dataset including Catalan and Chinese.
- `CPO CA-ZH`: Built by sourcing sentences from Catalan and Spanish Wikimedia resources, and subsequently directly translated to Chinese.

Below, we describe the data sourcing process for `CPO CA-ZH`, the translation methodology, and a more detailed composition of the two preference datasets.

3.1 Sourcing sentences

Original Source. Following the methodology of FLORES-200 , all source sentences were extracted from Wikimedia resources, which are publicly available under permissive licensing. To ensure that the selected data did not overlap with parallel datasets already included in the models, we verified that none of the chosen Wikimedia pages had corresponding versions in Chinese.

The dataset was divided into three (roughly) equal parts to ensure diversity and coverage across different domains. Approximately one-third of the sentences were collected from Catalan *Wikinews*[4], a collection of news articles, with content selected from ten distinct topics. These topics, chosen to maintain balance and variety, include science and technology, culture and leisure, law, economy, sports, environment, obituaries, politics, health, and incidents. The second portion of the dataset was drawn from Catalan *Wikipedia*, a general-purpose encyclopedia containing a wide range of topics. The final third was sourced from *Wikivoyage*, a travel guide featuring articles on travel tips, cuisine, and destinations worldwide. Since Catalan *Wikivoyage* is still under development and, as of January 3, 2025, contains only 31 articles, this portion was instead sourced from Spanish *Wikivoyage*, which is significantly more developed and includes 3,347 articles.

Sentence Selection. Sentences were selected using a systematic approach to ensure diversity. Articles were selected from each source domain by randomly generating URLs using the *requests* library.[5] Following the methodology of FLORES-200 , between 3 and 5 contiguous sentences were extracted from each selected article, avoiding very short or malformed sentences. For Catalan *Wikinews* and Catalan *Wikipedia*, sentences were chosen equally from the beginning, middle, and end of each article to capture varied contexts. For Spanish *Wikivoyage*, selected sentences represented different topics, such as "drinking and nightlife", "climate", "shopping", and "flora and fauna" (see Appendix B for detailed dataset statistics).

Each selected sentence was annotated with metadata, including the article ID, sentence ID, URL and topic. On average, 3.5 contiguous sentences were extracted per article, with URLs included to allow access to the full document, which can be useful for document-level translation.

3.2 Translation

We used GPT-4 (OpenAI et al., 2024), which has demonstrated performance comparable to junior translators (Yan et al., 2024), to translate Catalan sentences into Spanish and Chinese. For sentences sourced from Spanish *Wikivoyage*, GPT-4 was used to translate them into Catalan and Chinese (see Appendix A for the specific GPT-4 prompt). Given the linguistic similarities between Spanish and Catalan, high-quality translations are assumed for this pair. For the Chinese translations, a native Chinese-speaking translator conducted post-editing and revisions of the machine-translated sentences to ensure naturalness and accuracy.

3.3 Two preference datasets

The `CPO CA-ZH` dataset consists of 1,022 triplets. Each Catalan sentence sourced from Wikimedia

[4] https://ca.wikinews.org/wiki/Portada.

[5] For example, a GET request to https://es.wikivoyage.org/wiki/Especial:Aleatoria redirects to a random article on Spanish *Wikivoyage*.

served as the *prompt*. Machine-translated Chinese sentences from GPT-4 were used as the *rejected* translations, while human-revised translations were labeled as the *chosen* sentences. Although GPT-4 translations are of relatively high quality, the goal of CPO is to train the model to recognize and prefer human-revised translations, thereby aligning more closely with human preferences.

In `CPO FLORES DEV`, there are in total 997 triplets. Catalan sentences from the FLORES-200 *dev* split served as the *prompt*. GPT-4 was used to translate the original English sentences into Chinese, producing the *rejected* translations. The original Chinese translations from the FLORES-200 *dev* split, which were also translated from English sentences, served as the *chosen* sentences.

In summary, `CPO CA-ZH` features direct Catalan-to-Chinese translations, while `CPO FLORES DEV` relies on an English pivot for generating Chinese translations. Both datasets use GPT-4 generated translations as *rejected* outputs and human-revised or human-produced translations as *chosen* outputs.

4 Entities in the CPO CA-ZH dataset

To analyze the key entities discussed in our `CPO CA-ZH` dataset, we used *spaCy* (version 3.8.3) to extract proper noun phrases and their corresponding frequencies from the Catalan sentences. These entities were then compared with those in the FLORES-200 and NTREX (Federmann et al., 2022) to assess how the topics in our dataset differ from those in existing datasets.

Overall, the `CPO CA-ZH` dataset is more focused on geographically specific topics, with frequent references to entities such as *Barcelona*, *Espanya* (Spain), and *Catalunya* (Catalonia). These entities are either absent or significantly less prominent in the other datasets. In contrast, the *dev* and *devtest* splits of the FLORES-200 prominently feature *Estats Units* (United States) as the most frequent entity, and the NTREX also tends to focus more often on entities like *Trump* and *USA*. For a complete comparison, see Table 1 and the frequency of each phrase in Appendix C.

This analysis suggests that our `CPO CA-ZH` dataset is more localized and culturally specific, emphasizing topics relevant to the region, whereas the FLORES-200 and NTREX are more focused on the United States and globally oriented topics.

5 Experiments

We applied CPO to the M2M-100-1.2B model using each of the two preference datasets introduced in Section 3.3. To assess the models after CPO training, we evaluated their translation performance on the FLORES-200 *devtest* split, which primarily focuses on topics relevant to the United States and global contexts. In addition, we conducted A/B testing on translations of 100 sentences containing localized terms specific to Catalonia. This allowed us to evaluate and compare the models' capabilities in handling more region-specific translations.

5.1 Training setup

We used the `facebook/m2m100_1.2B` (Fan et al., 2021)[6], a seq-to-seq model trained for multilingual translation, as the base model. It covers 100 languages, including Catalan and Mandarin Chinese.

Fine-tuning was performed using the Hugging Face's CPOTrainer class[7] which is compatible with the M2M-100 encoder-decoder architecture. We adhere to the default β value of 0.1 as suggested by Rafailov et al. (2024). The fine-tuning process involved a total batch size of 5, training for 6 epochs. The learning rate started at 8e-6 and linearly decayed throughout training. Checkpoints were saved every 50 steps and evaluated on the FLORES-200 *devtest* set. Training was conducted on a single NVIDIA H100 GPU with 64GB of RAM and completed in approximately 10 minutes.

5.2 Inference

Inference for all models was conducted using beam search with a beam size of 5, limiting the translation length to 200 tokens.

6 Results

6.1 Evaluation on FLORES devtest

This section reports the evaluation results of the models on the FLORES-200 *devtest* split for the Catalan→Chinese translation direction. The evaluation was conducted using `MT Lens` (Gilabert et al., 2024).[8] To provide a comprehensive assessment, we report a variety

[6]The smaller M2M-100-418M model often generates unknown tokens when translating from Catalan to Chinese (e.g., unknown tokens appear in 15% of translations on the FLORES-200 *devtest* split). To better support our evaluation of the translation of localized terms, we chose the larger 1.2B model, which provides greater vocabulary coverage for our experiments.

[7]https://huggingface.co/docs/trl/en/cpo_trainer

[8]

Our Dataset	FLORES-200 Dev	FLORES-200 Devtest	NTREX
Barcelona	Estats Units (United States)	Estats Units (United States)	Trump
Estats Units (United States)	Terra (Earth)	Terra (Earth)	EUA (USA)
Europa (Europe)	Xina (China)	Austràlia (Australia)	Regne Unit (United Kingdom)
Universitat (University)	EUA (USA)	Alemanya (Germany)	Xina (China)
Espanya (Spain)	Europa (Europe)	França (France)	Kavanaugh
Catalunya (Catalonia)	Àfrica (Africa)	Japó (Japan)	Corea (Korea)
Xina (China)	Sol (Sun)	Europa (Europe)	Palu
França (France)	Itàlia (Italy)	Hong Kong	Nord (North)
Madrid	Alemanya (Germany)	Taiwan	UE (EU)
Alemanya (Germany)	Turquia (Turkey)	Suècia (Sweden)	Washington

Table 1: Top 10 most frequent proper noun phrases across datasets

of metrics: BLEU (version 2.3.1), BLEURT (`lucadiliello/BLEURT-20-D12`), COMET (`Unbabel/wmt22-comet-da`), COMET-Kiwi (`Unbabel/wmt22-cometkiwi-da`), MetricX (`google/metricx-23-xl-v2p0`), MetricX-QE (`google/metricx-23-qe-xl-v2p0`) and statistical significance testing using paired bootstrap resampling (Koehn, 2004).

As shown in Table 2, BLEU scores (Papineni et al., 2002) indicate that `+CPO FLORES DEV` achieved a significant improvement in n-gram overlap between model translations and the reference, while the improvement with `+CPO CA-ZH` was not statistically significant. This result was expected, given the similarities between FLORES-200 *dev* (used for training) and FLORES-200 *devtest*.

In contrast, neural-based metrics such as BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020), and MetricX (Juraska et al., 2023), as well as neural-based reference-free metrics like COMET-Kiwi (Rei et al., 2022) and MetricX-QE, suggest that `+CPO CA-ZH` led to greater improvements in translation quality. This indicates that `+CPO CA-ZH` improves aspects of translation quality such as semantic accuracy and fluency, without necessarily relying on the same n-gram phrases as the reference translation.

6.2 Evaluation on more culture-specific entities and data

In addition to the FLORES-200 *devtest* set, we assessed the models on sentences that contain Catalan- and Spanish-specific topics and culturally significant entities. We randomly selected 100 sentences from the Catalan Entity Identification and Linking dataset (Gonzalez-Agirre et al., 2024)[9] and ensured that most selected sentences contained regionally or culturally specific entities.

We used the two fine-tuned models to generate Chinese translations of these sentences. The translations were then assessed through A/B testing by two annotators: a linguist (the author) fluent in both Catalan and Chinese, and a professional Catalan-Chinese translator with eight years of experience. The annotators evaluated which translation more accurately conveyed the original meaning and sounded more natural. To measure the consistency between the annotators' preferences, we calculated the inter-annotator agreement using Cohen's kappa statistic with the *sklearn* library (version 1.5.2). The kappa score was 0.68, indicating substantial agreement according to the guidelines by Landis and Koch (1977).

Translations produced by `+CPO CA-ZH` were preferred more often (Annotator 1: 59% of the time; Annotator 2: 68%) compared to `+CPO FLORES DEV`. Among the 85 items where both annotators agreed, 56 (66%) favored `+CPO CA-ZH`. These results indicate a general preference for the translations from `+CPO CA-ZH`.

Furthermore, through manual examination, `+CPO CA-ZH` produced more accurate translations for region-specific terms and exhibited better transliteration capabilities from Catalan to Chinese. Examples of these translations are shown in Table 3, with the complete translated sentences available in the Appendix D. Even though these terms have never appeared in our preference dataset, aligning the model with localized data improved its ability to accurately translate and transliterate region-specific terminology. This highlights the effectiveness of incorporating culturally and regionally relevant data into the training process for practical use.

[9]This datset comprises sentences from tweets, news articles, reports, forums, encyclopedias, parliamentary proceedings, and reviews, and was originally designed for Named Entity Recognition.

Models	BLEU ↑	BLEURT ↑	COMET ↑	COMET-Kiwi ↑	MetricX ↓	MetricX-QE ↓
M2M100 1.2B	28.23	0.65	0.82	0.77	3.12	2.71
+ CPO CA-ZH	29.15	0.68	0.84 *	0.79 * †	2.51 * †	1.91 * †
+ CPO FLORES DEV	29.58 * †	0.67	0.84 *	0.77	2.60 *	2.15 *

* Significant improvement over the baseline M2M100 1.2B ($p < 0.05$).
† Significant difference between the two CPO-tuned models ($p < 0.05$).
Note: Significance testing was not performed for BLEURT as it is currently unsupported by `MT Lens`.

Table 2: The results in CA→ZH for FLORES-200 devtest set.

Catalan phrase in sentences	**Explanation**	**+ CPO FLORES DEV**	**+ CPO CA-ZH**
Bàsquet Girona	professional basketball club based in Girona	吉罗纳篮球队(Girona Basketball Team)	吉罗纳篮球俱乐部(Girona Basketball Club)
autònoms	self-employed workers or freelancers	自治人(Autonomous People)	自主经营者(Self-Employed)
Sant Feliu de Llobregat	municipality in the province of Barcelona	罗布拉格(Robrag)	圣费利乌·德·卢布雷加特(Sant Feliu de Llobregat)
Blanes	municipality in Catalonia	布莱斯(bù lái sī)	布拉内斯(bù lā nèi sī)
Corredor Mediterrani	Mediterranean Corridor, a major rail transport network in Europe	地中海跑道(Mediterranean Track)	地中海走廊(Mediterranean Corridor)
merder dels okupes	the mess caused by squatters; colloquial	混乱(Chaos)	占领活动的混乱(Chaos of Squatter Activities)

Table 3: Examples of Chinese translation of Catalan and Spanish region-specific terms, with English translations or *pinyin* provided in parentheses.

7 Conclusion

Many existing machine translation benchmarks, such as FLORES-200, rely on English as a pivot language for non-English language pairs. This approach can overlook the linguistic and cultural specificity of direct translations, particularly for language pairs like Catalan-Chinese (CA-ZH), where structural differences, idiomatic expressions, and cultural references may not have direct equivalents in English. To address this gap, we present a CA-ZH parallel dataset containing 1,022 sentences sourced from Catalan and Spanish Wikimedia and directly translated into Mandarin Chinese. Unlike most existing benchmarks, our dataset prioritizes linguistic and cultural authenticity by capturing regional nuances specific to Catalonia and Spain. This localization ensures that translations reflect real-world usage rather than being filtered through a more globalized or English-centric lens. By comparing our dataset to the FLORES-200 *dev* set, we demonstrate the benefits of aligning machine translation (MT) systems with culturally and regionally grounded data. This direct translation approach outperforms English-pivoted methods, which often introduce biases from the English-speaking world. Additionally, our dataset enables more accurate pronunciation mapping and transliteration between Catalan and Chinese, further improving transliteration quality for practical applications. Our work highlights the importance of developing non-English-centric datasets to better serve low-resource language pairs. We hope that the release of this dataset will encourage further research into localized, culturally rich resources and improve MT systems for real-world use.

Limitations

One limitation of our dataset is its relatively small size. While we aimed to create a high-quality

dataset, the process of finding linguists and professional translators who are fluent in both Chinese and Catalan, as well as knowledgeable about Catalan culture, is costly. However, this constraint also ensures that the dataset (1,022 sentences) allows for meaningful comparisons with the FLORES dev set (997 sentences), maintaining fairness in evaluation.

That said, the limited number of sentences, combined with the fact that we did not explicitly ensure that every randomly selected document discusses Catalan or Spanish culture during the sentence sourcing process, means that the dataset could have been richer in regionally and culturally specific topics. Future expansions could address this by incorporating more diverse sources that better reflect the cultural and linguistic nuances of Catalan-speaking communities.

Ethical statements

The annotators were fairly compensated at a rate of approximately 20 euros per hour, ensuring ethical payment for their work. Sentences in our dataset were sourced from Wikimedia under a public license, adhering to open data principles and respecting intellectual property rights.

Carbon impact Statement

This work considers the environmental impact of computational resources used in model training. Each CPO training runs in 10min on 1 NVIDIA H100 GPU, and draws 201.47 Wh. Based in Spain, this has a carbon footprint of 34.46 g CO_2e, which is equivalent to 3.76e-02 tree-months, (calculated using green-algorithms.org v2.2 (Lannelongue et al., 2021)). Compared to large-scale deep learning methods, which can emit several metric tons of CO_2e, our approach remains computationally efficient and environmentally sustainable. In fact, the emissions per run are comparable to just a few Google searches, highlighting the low-carbon footprint of this training process while maintaining high model performance.

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 DeepR3 (TED2021-130295B-C32) funded by MCIN/AEI/10.13039/501100011033 and European Union NextGeneration EU/PRTR.

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

Yongjian Chen, Antonio Toral, Zhijian Li, and Mireia Farrús. 2024. Improving NMT from a low-resource source language: A use case from Catalan to Chinese via Spanish. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 229–245, Sheffield, UK. European Association for Machine Translation (EAMT).

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

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.

Marta R Costa-Jussa, Noé Casas, Carlos Escolano, and José AR Fonollosa. 2019. Chinese-Catalan: A neural machine translation approach based on pivoting and attention mechanisms. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(4):1–8.

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*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

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.

Javier García Gilabert, Carlos Escolano, Audrey Mash, Xixian Liao, and Maite Melero. 2024. Mt-lens: An all-in-one toolkit for better machine translation evaluation. *arXiv preprint arXiv:2412.11615*.

Aitor Gonzalez-Agirre, Montserrat Marimon, Carlos Rodriguez-Penagos, Javier Aula-Blasco, Irene Baucells, Carme Armentano-Oller, Jorge Palomar-Giner, Baybars Kulebi, and Marta Villegas. 2024. Building a data infrastructure for a mid-resource language: The case of Catalan. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2556–2566, Torino, Italia. ELRA and ICCL.

Tianxiang Hu, Haoxiang Sun, Ruize Gao, Jialong Tang, Pei Zhang, Baosong Yang, and Rui Wang. 2024. Sjtu system description for the wmt24 low-resource languages of spain task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 943–948.

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. 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.

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.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: quantifying the carbon footprint of computation. *Advanced science*, 8(12):2100707.

Zixuan Liu. 2022. Improving Chinese-Catalan machine translation with Wikipedia parallel corpus. Master's thesis, Universitat Pompeu Fabra, Barcelona.

Palanichamy Naveen and Pavel Trojovskỳ. 2024. Overview and challenges of machine translation for contextually appropriate translations. *Iscience*, 27(10).

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, 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, Ben 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, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha 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, Johannes 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, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz 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, Mateusz 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 Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ry-

der, Mario 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 Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll 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, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

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.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, 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. *arXiv preprint arXiv:2401.08417*.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *Preprint*, arXiv:2407.03658.

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, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

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.

Chenyue Zhou. 2022. Building a catalan-chinese parallel corpus from wikipedia for use in machine translation. Master's thesis, Universitat Pompeu Fabra, Barcelona.

A Prompt for translations

Adhering to the prompt format for translation as utilized by Xu et al. (2024) for GPT models, we use the same prompt for GPT-4 in our study, as shown in Figure 1.

GPT-4 Prompt

System:
You are a helpful translator and only output the result.

User:
Translate this from <source language> to <target language>, <source language>:
<source sentence>
<target language>:

Figure 1: The prompt employed for GPT-4 to perform translations.

B Statistics of the CPO CA-ZH dataset

The CPO CA-ZH dataset includes sentences collected from three primary sources: Catalan Wikinews, Catalan Wikipedia, and Spanish Wikivoyage. Approximately one-third of the sentences come from each source:

Source	Wikinews	Wikipedia	Wikivoyage
n. sent	328	341	353

Table 4: Number of sentences collected from different sources.

For Catalan Wikinews and Catalan Wikipedia, sentences were chosen roughly equally from the beginning, middle, and end of each article to capture varied contexts:

Sentence Position	**Count**
Middle	231
End	222
Beginning	216

Table 5: Distribution of sentence positions for Catalan Wikinews and Catalan Wikipedia.

The statistics for the Wikinews portion of the dataset are shown in Table 6. The topics, along with their English translations, are as follows: Ciència i Tecnologia (Science and technology), Cultura i esplai (Culture and leisure), Dret (Law), Economia (Economy), Esports (Sports), Medi ambient (Environment), Necrologia (Obituaries), Política (Politics), Salut (Health), Successos (Incidents).

C Top 10 most frequent proper noun phrases across datasets

Table 7 shows the top 10 most frequent proper noun phrases and their frequency in our CPO CA-ZH dataset, FLORES-200 *dev* split, FLORES-200 *devtest* split, and the NTREX dataset.

D Examples of translation of region-specific terms

In Section 6.2, we have only shown translation of the phrases. Table 8 below shows the translation of the full sentences where these phrases come from.

Topic	# Articles	# Sentences
Ciència i tecnologia	9	29
Cultura i esplai	9	28
Dret	10	36
Economia	10	35
Esports	10	35
Medi ambient	10	35
Necrologia	9	28
Política	10	37
Salut	10	34
Successos	10	31
Total	97	328

Table 6: Statistics of the Wikinews portion of the dataset.

Our Dataset	Freq.	FLORES-200 Dev	Freq.	FLORES-200 Devtest	Freq.	NTREX	Freq.
Barcelona	18	Estats Units	14	Estats Units	17	Trump	64
Estats Units	13	Terra	10	Terra	11	EUA	61
Europa	12	Xina	9	Austràlia	9	Regne Unit	42
Universitat	9	EUA	8	Alemanya	9	Xina	40
Espanya	9	Europa	7	França	8	Kavanaugh	36
Catalunya	8	Àfrica	7	Japó	7	Corea	35
Xina	7	Sol	7	Europa	7	Palu	29
França	7	Itàlia	6	Hong Kong	6	Nord	27
Madrid	6	Alemanya	6	Taiwan	6	UE	26
Alemanya	6	Turquia	6	Suècia	6	Washington	22

Table 7: Top 10 most frequent proper noun phrases and their frequency across datasets

Catalan sentence (keyword)	+ CPO FLORES DEV	+ CPO CA-ZH
Són també representatius el **Bàsquet Girona** (fundat el 2014 per Marc Gasol), actualment a l'ACB, La Salle Girona, el Vedruna Girona, el Sant Narcís, el Club Bàsquet Onyar o el CESET	吉罗纳篮球队(由马克·加索尔(Marc Gasol)于2014年创立)也代表了这支球队,目前属于ACB的La Salle Girona、维德鲁纳·吉罗纳(Vedruna Girona)、圣纳基斯(San Narcís)、奥尼亚尔篮球俱乐部(Club Bàsquet Onyar)和CESET。	吉罗纳篮球俱乐部(由马克·加索尔于2014年创办)也代表着它,目前它属于ACB、拉萨尔·吉罗纳俱乐部、维德鲁纳·吉罗纳俱乐部、圣纳西斯俱乐部、奥尼亚尔篮球俱乐部以及塞塞斯特俱乐部。
Més de 6.500 **autònoms** sol·liciten l'ajut del Govern i Treball estudia ampliar-lo i fer-lo complementari amb l'estatal.	超过6500名自治人向政府申请了援助,劳工正在考虑扩展援助,并将该援助与州政府互补。	超过6500名自主经营者申请了政府的援助,劳工部门正在考虑扩大援助计划并将其与州政府的援助互补。
Les obres per soterrar les vies a **Sant Feliu de Llobregat** començaran aquest 2019.	罗布拉格(Sant Feliu de Llobregat)的道路埋葬工作将于2019年开始。	圣费利乌·德·卢布雷加特(Sant Feliu de Llobregat)的道路埋葬工程将于2019年开始。
Per al passeig Marítim de la **ciutat de Blanes** se li va encarregar el monument homenatge a l'escriptor Joaquim Ruyra (1958) amb ocasió del centenari del seu naixement, ho va representar com el frare caputxí «sant Sadurní de Croïlles» protagonista de la seva obra Les coses benignes, la intenció de l'escultor va ser expressar la gran unió que existeix entre la literatura de Ruyra i l'esperit franciscà.	对于布莱斯(Blanes)市的海上漫游活动,他在1958年为作者约阿基姆·鲁伊拉(Joaquim Ruyra)诞生100周年纪念日提供纪念碑,他将其描绘成"克鲁伊莱斯圣萨德尼"(Saint Sadurní de Croïlles),其著作《善良的事物》(The Good Things)的主角。雕塑家意图是表达鲁伊拉(Ruyra)文学与法兰西斯(Franciscan)精神之间存在的巨大联系。	对于布拉内斯市的海上散步,在1958年出生百周年之际,他委托了作家约阿基姆·鲁伊拉(Joaquim Ruyra)的纪念碑。他将其描绘成"圣萨杜尼·德·克鲁伊莱斯"(Saint Sadurní de Croïlles)作为他的作品主角。雕塑家的目的是表达鲁伊拉文学与弗朗西斯克精神之间深远的联系。
De la Serna assenyala que el ritme inversor al **Corredor Mediterrani** no depèn d'un problema econòmic sinó administratiu.	德尔·塞尔纳(De la Serna)指出,地中海跑道的变速不是因为经济问题,而是因为行政问题。	德拉·塞尔纳指出,地中海走廊的投资步伐并非因为经济问题,而是因为行政问题。
Aquest **merder dels okupes** a Barcelona i en extensió a tota Catalunya va ser propiciat per la Colau, que oblidem molt ràpid les coses.	巴塞罗那的这一混乱,以及整个加泰罗尼亚的混乱,是由"科洛"(La Colau)造成的,我们很快就忘了这些事情。	这场在巴塞罗那以及整个加泰罗尼亚的占领活动的混乱,是由劳拉·科劳(La Colau)推动的。我们很快就忘记了这些事情。

Table 8: Examples of translation of Catalan and Spanish region-specific terms in sentences.

Instruction-tuned Large Language Models for Machine Translation in the Medical Domain

Miguel Rios
Centre for Translation Studies, University of Vienna
miguel.angel.rios.gaona@univie.ac.at

Abstract

Large Language Models (LLMs) have shown promising results in machine translation, particularly for high-resource settings. However, in specialised domains such as medicine, their translation quality underperforms compared to standard Neural Machine Translation models, particularly regarding terminology consistency. In this study, we investigate the impact of instruction tuning for enhancing LLM performance in machine translation for the medical domain. We compare baseline LLMs with instruction-tuned models, and explore the impact of incorporating specialised medical terminology into instruction-formatted fine-tuning datasets. Our results show that instruction tuning significantly improves LLM performance according to automatic metrics. Furthermore, error analysis based on automatic annotation shows a substantial reduction in translation errors in the instruction-tuned models compared to the baselines.

1 Introduction

Current state-of-the-art Large Language Models have shown promising results in machine translation for high-resource language pairs and domains (Bawden and Yvon, 2023). However, in low-resource domains (e.g. medical) LLMs have shown lower performance compared to standard neural machine translation (NMT) models (Bawden and Yvon, 2023; Pourkamali and Sharifi, 2024). The accuracy and consistency in the machine translation of terminology, syntax, and document structure is crucial for users, researchers, and translators who post-edit machine translated documents in high-risk domains (Almahasees et al., 2021; Pang et al., 2024). Moreover, the introduction of in-domain translation constraints during generation into neural models is currently an open problem (Saunders et al., 2019; Alves et al., 2023; Hauhio and Friberg, 2024).

LLMs are trained to perform different Natural Language Processing (NLP) tasks such as summarisation, question answering, and translation, where users interact with the models via instructions (e.g. chat interface) (Touvron et al., 2023; OpenAI et al., 2024; Dubey et al., 2024). Instruction-tuning is a technique that leverages datasets from different NLP tasks, structured as prompts, for fine-tuning LLMs to enhance generalisation across novel tasks and domains (Chung et al., 2022). For example, in machine translation (MT), translating a segment from the European Medicines Agency corpus with specialised terminology the prompt can be framed as: "*Glossary: medicine -> medicamento.*
Translate the source text from English to Spanish following the provided translation glossaries.
English: The medicine was effective in patients with all three types of homocystinuria.
Spanish: ".

Moreover, Alves et al. (2024) instruction-tuned Llama-2 (Touvron et al., 2023) to perform translation related tasks, such as segment and document level translation, post-editing, terminology aware translation, and error annotation. The controlled generation of MT output with the correct terminology, segment length, or syntax can be framed as an instruction-tuning task for LLMs. Thus, improving the workflow of translation during post-editing with an instruction-following (i.e. chat) interface for an LLM tuned on a specific domain.

We seek to answer the following research question: Does instruction-tuning based on terminology rules improve translation quality on LLMs? In this paper, we show results for adding specialised medical dictionaries into fine-tuning for LLMs. In particular, we follow the methodology from (Alves et al., 2024) by incorporating terminology information into the instruction-tuning datasets. Unlike (Alves et al., 2024), our approach relies on openly

Proceedings of Machine Translation Summit XX Volume 1, pages 162–172, June 23-27, 2025

available medical dictionaries and employs simple heuristics to construct instruction-tuning datasets. An instruction-based interface could facilitate the interaction between professional translators and LLMs, and enables model customisation via the integration with user-defined terminology dictionaries.

Our contributions are as follows: We use parameter-efficient fine-tuning (PEFT) and quantisation of large language models (LLMs) for in-domain translation. We leverage medical dictionary term pairs with parallel data to construct prompts that guide LLMs in translating specific terminology.

We evaluate FLAN-T5 (Chung et al., 2022), Llama-3 (Dubey et al., 2024), and Tower (Alves et al., 2024) for English-Spanish, English-German, and English-Romanian language pairs in a split of a medical domain dataset.

The instruction-tuned models outperformed the baseline models with the automatic metrics BLEU (Papineni et al., 2002), chrF (Popović, 2015), and COMET (Rei et al., 2020). Moreover, instruction-tuning improves the overall accuracy of the terminology. Finally, we evaluate the two best models with automatic error annotation (Guerreiro et al., 2024), and quality estimation (Rei et al., 2023).

2 Background and Related Work

Auto-regressive language models predict the next token in a sequence given a prefix context (Jelinek, 1998; Bengio et al., 2000), where LLMs are pre-trained with large amounts of texts followed by fine-tuning on different downstream tasks (OpenAI et al., 2024). In addition, Chung et al. (2022) propose to fine-tune LLMs with a mixture of several NLP datasets into an instruction format to improve: generalisation to unseen tasks, and generation given instruction prompts. For a machine translation task, the LLM is conditioned on a user defined prompt that consists of a translation instruction along with the source text to be translated (Pang et al., 2024). During testing, zero-shot prompting involves querying an LLM with a test input that was not present in the training data. For example, MT instruction includes a prompt asking to translate from a source language to a target language, and the corresponding source text. However, few-shot prompting provides a few examples of the translation task along with the test input to guide the LLM generation. In MT, few-shot examples consist of parallel source and human-translated sentences.

Supervised fine-tuning (SFT) is one of the most popular techniques for domain adaptation in LLMs, where models continue their training with a sample of in-domain data (Alves et al., 2023; Eschbach-Dymanus et al., 2024). However, SFT for LLMs requires large amounts of computational resources, given that during training models update billions of parameters. The goal of Parameter-efficient fine-tuning (PEFT) is to update (i.e. tune) a minimal set of parameters to achieve a similar performance compared to full SFT on downstream tasks. Hu et al. (2021) propose low-Rank adaptation (LoRA) that freezes all the pre-trained model parameters and adds adapter trainable low-rank decomposition matrices of parameters into each layer of the model.

Moreover, Dettmers et al. (2023) propose that during fine-tuning to quantise the parameters of the pre-trained model into fewer bits (e.g. 4-bit) and keep the LoRA adapters with standard precision, thus reducing the memory usage. PEFT and quantisation with QLoRA enables academic translation practitioners to fine-tune LLMs with limited computing resources. Llama versions 2 and 3 (Touvron et al., 2023; Dubey et al., 2024) are open-source LLMs with different parameter scales, which are instruction-tuned for multiple Natural Language Processing tasks. Moreover, Llama has become the base model for the MT related work (Alves et al., 2023; Pang et al., 2024; Eschbach-Dymanus et al., 2024).

Zhang et al. (2023) compared 15 baseline LLMs and fine-tuned with QLoRA on different MT tasks (e.g. segment and document level translation) for the French-English language pair. Llama-2 outperformed other LLMs, fine-tuning improves performance on models that struggle on a few-shot setup, and QLoRA is potentially superior to full SFT in terms of efficiency. Alves et al. (2023) compared instruction tuning with LoRA to few-shot prompting using Llama-2 in various language pairs. Fine-tuning outperforms the few-shot learning, is comparable to full SFT, requires few training data, and tackles over generation. However, LLMs struggle with translation directions out of English (en-xx). Alves et al. (2024) proposed Tower with a focus on translation related tasks, for example, document level translation, post-editing, and terminology-aware prompts. Tower is based on the continued training of Llama-2 with parallel translation data, and is followed by instruction-tuning for

the MT tasks.

Zheng et al. (2024) proposed to fine-tune LLMs based on prompts, and compared it to LoRA for domain adaptation in IT for Chinese-English and English-Chinese MT. Moreover, Zheng et al. (2024) incorporate IT terminology by few-shot prompting and chain-of-thought. The template used for the proposed prompt-tuning model has a substantial impact on performance, and the introduction of terminology with simple prompt rephrasing outperforms chain-of-thought. Eschbach-Dymanus et al. (2024) studied domain adaptation for business IT with LLMs. They compared full SFT, LoRA, different prompting techniques, and standard NMT. Finally, Eschbach-Dymanus et al. (2024) defined guidelines for domain adaptation with LLMs. Moslem et al. (2023) evaluate LLMs for translation on specialised domains (e.g. medical COVID-19), and incorporate terms from glossaries into their prompts to tackle issues with no retrieved matches in few-shot learning. Jerpelea et al. (2025) developed a parallel dataset for the low-resource languages Romanian, and Aromanian, they instruction-tuned Llama-3 for Romanian, and compared multilingual NMT, GPT, Llama-3, and Tower for translation.

We followed (Alves et al., 2023, 2024) for our experimental design. Unlike previous work on LLM for MT, our approach focuses on the medical domain, relies on openly available medical dictionaries, employs simple heuristics to construct the instruction-tuning datasets, and uses efficient tuning techniques. In particular, we evaluate the impact of instruction tuning for improving terminology translation in LLMs.

3 Experimental Setup

3.1 Data

We use the corpus of the European Medicines Agency (EMEA) (elr) for the English-Spanish (en-es), English-German (en-de), and English-Romanian (en-ro) language pairs. The EMEA corpus contains multilingual PDF documents from the European Medicines Agency, automatically converted to text and aligned at the segment level. We randomly split the EMEA corpus into 20K segments for each language pair. These subsets were then merged into a single tuning dataset of 60K segments. Furthermore, we created separate validation and test sets, each containing 500 segments per language pair.

3.2 Terminology Annotation

The Interactive Terminology for Europe (IATE)[1] is a terminology management system from EU institutions that covers different domains (e.g. economics, law, health). For our source and target language pairs, we downloaded the IATE database in the *health* domain (*id 2841*). We only used terms with quality 3 (reliable) and 4 (very reliable) stars (human annotated quality scores), resulting in 38,898 terms for en-es, 49,828 terms for en-de, and 9,551 terms for en-ro.

We incorporate medical terms as translation instructions by identifying term pairs within each aligned segment. For every aligned segment, we retrieve candidate terms using strict matching, which requires the presence of a candidate pair in both the source and target segments. If one or more candidate pairs are identified, we include them in the instruction template within the prompt. For example, an instruction-tuning input in en-es "*spectrum of activity -> espectro de actividad, amoxicillin -> amoxicilina, and activity -> actividad*" are term pairs identified in the parallel segment:

```
Glossaries:
"spectrum of activity" -> "espectro de
    actividad"
"amoxicillin" -> "amoxicilina"
"activity" -> "actividad"
Translate the source text from English
    to Spanish following the provided
    translation glossaries.
English: Amoxicillin is susceptible to
    degradation by beta-lactamases
    produced by resistant bacteria and
    therefore the spectrum of activity
    of amoxicillin alone does not
    include organisms which produce
    these enzymes.
Spanish: La amoxicilina es sensible a
    la degradación por las beta-
    lactamasas producidas por bacterias
     resistentes y por tanto el
    espectro de actividad de la
    amoxicilina sola no incluye
    microorganismos productores de
    estas enzimas.
```

When no candidates are identified within a segment, the instruction consists only of the translation task prompt. For example, an instruction-tuning input in en-es:

```
Translate the source text from English
    to Spanish.
English: Do not use Cymevene if you are
    breast-feeding.
Spanish: No use Cymevene si está en
    periodo de lactancia.
```

[1] https://iate.europa.eu/download-iate

The prompt templates for the baseline models are defined in Section B, and we perform zero-shot prompting to generate translations for en-es, en-de, and en-ro. For example, a test input in en-es:

```
Glossary:
"insulin" -> "insulina"
Translate the source text from English
    to Spanish following the provided
    translation glossaries.
English: Within-subject variability of
    the time action profile of Levemir
    and NPH insulin Pharmacodynamic
    Endpoint
Spanish:
```

3.3 Models

We use the HuggingFace transformers framework for the baseline LLMs (Wolf et al., 2020), and PEFT (Mangrulkar et al., 2022) for the instruction-tuning with QLoRA. Our baseline LLMs are as follows: FLAN-T5-large[2] (783M parameters), an encoder-decoder model; Llama-3-8B[3], an instruction-tuned LLM for NLP tasks; and Tower-7B[4], an instruction-tuned LLM for MT tasks. We evaluated two distinct instruction-tuned model architectures: encoder-decoder model based on FLAN-T5, and auto-regressive LLMs based on Llama.

We use QLoRA with a 4-bit quantisation to fine-tune each baseline model for one epoch on the tuning dataset (60K segments). The values for QLoRA and tuning hyper-parameters for each model are defined in Section A, for FLAN-T5 7, Llama-3-8B 8, and Tower-7B 9. Finally, for generation, we use zero-shot prompting and stochastic decoding with top-p sampling $p = 0.9$. We release our scripts and data on GitHub at: https://github.com/HAITrans-lab/instruction-tuned-medical-LLM

4 Results

We show results with automatic metrics and terminology accuracy for FLAN-T5, Llama-3 and Tower for the en-es, en-de and en-ro language pairs. Moreover, we show automatic error annotation, and quality estimation scores for the best performing models.

4.1 Automatic Metrics

We evaluated all models with BLEU, chrF, and COMET in the test split. Table 1 shows the comparison between the baselines and the instruction-tuned models with QLoRA. The BLEU, chrF, and COMET scores for the instruction-tuned models are statistically significant ($p < 0.05$) for all models.

To prevent over-generation and improve the performance of Llama-3, we post-processed the output by cutting it at the first appearance of the end-of-sequence token "<|*eot_id*|>". As noted by Zhang et al. (2023), Llama models repeat the translation output or produce *assistant* suggestions to improve the prompts along with the translation.

In Table 1, Tower and the QLoRA Tower outperform the other models with the automatic metrics for en-es, and en-de. However, Romanian (en-ro) is not present in the original Tower fine-tuning for MT. Tower is based on LLaMA-2 which is not focused on multilingual data, in contrast to Llama-3. Moreover, QLoRA tuning produced improvements for all models.

As shown in Table 1, Tower and QLoRA Tower achieved the highest automatic metric scores for en-es, and en-de. However, the original Tower model was not fine-tuned for en-ro MT. Furthermore, Tower is based on LLaMA-2, which is less focused on multilingual data compared to Llama-3. Nonetheless, the QLoRA models consistently improved performance across all models. The bold numbers are the best automatic scores across all models for a given language pair.

Terminology Accuracy We compute the accuracy of the terminology in the MT output compared to the reference translations. To compute accuracy, the exact term must be present in both the MT segment and the database to be correct. Table 2 shows the accuracy scores for the terminology. Instruction-tuning improves the accuracy of terms across models, where Flan-T5 followed by Tower achieve the highest terminology accuracy performance. We observed that the LLM produced translations with increased terminology accuracy for the high-resource language pairs, en-de and en-es.

4.2 Automatic Error Annotation

Automatic metrics are not designed to identify specific translation errors in MT outputs, for example, errors in terminology. Multidimensional Quality Metrics (MQM) (Lommel et al., 2014) are based on manually classifying and annotating errors using predefined categories. The MQM error typology

[2] google/flan-t5-large
[3] meta-llama/Meta-Llama-3.1-8B-Instruct
[4] Unbabel/TowerInstruct-7B-v0.2

Model	en-es↑			en-de↑			en-ro↑		
	BLEU	chrF	COMET	BLEU	chrF	COMET	BLEU	chrF	COMET
FLAN-T5	28.51	57.11	0.73	14.76	43.86	0.63	17.34	45.00	0.64
QLoRA FLAN-T5	36.43	63.40	0.78	25.45	54.93	0.72	28.65	57.44	0.77
Llama-3-8B	34.07	63.02	0.79	25.44	58.08	0.78	24.99	53.17	0.76
QLoRA Llama-3-8B	45.07	67.74	0.85	36.30	62.21	0.84	**35.97**	**61.19**	**0.85**
Tower-7B	42.27	66.31	0.86	34.80	62.45	0.85	18.20	44.86	0.69
QLoRA Tower-7B	**48.88**	**70.36**	**0.87**	**42.11**	**67.62**	**0.87**	23.93	50.57	0.78

Table 1: Comparing the baseline and QLoRA fine-tuned LLMs with **automatic metrics** for the en-es, en-de, and en-ro language pairs.

Model	en-es↑	en-de↑	en-ro↑
FLAN-T5	0.72	0.45	0.38
QLoRA FLAN-T5	0.90	**0.91**	**0.90**
Llama-3-8B	0.59	0.53	0.44
QLoRA Llama-3-8B	0.69	0.68	0.51
Tower-7B	0.88	0.79	0.58
QLoRA Tower-7B	**0.91**	0.86	0.68

Table 2: Comparing the baseline and QLoRA fine-tuned LLMs with **terminology accuracy** for the en-es, en-de, and en-ro language pairs.

covers high-level error categories (e.g. Accuracy, linguistic conventions, style, etc.), where each category can be further expanded into fine-grained categories (e.g. Accuracy into Mistranslation, addition, untranslated, etc.). Expert translators identify an error in the MT output, label it with a category from the typology, and also assign a severity score to it. The severity weights defined in (Freitag et al., 2021) are: minor $\times$ 1 (MIN), major $\times$ 5 (MAJOR), and critical $\times$ 10 (CRIT). The MQM score is defined as follows:

$$\text{MQM} = 100 \cdot \left(1 - \frac{10 \cdot \text{critical} + 5 \cdot \text{major} + \text{minor}}{\text{tokens}}\right), \quad (1)$$

We use XCOMET (Guerreiro et al., 2024) to produce automatic MQM annotations. XCOMET only annotates the error spans in the MT output with severities [5], and the corresponding prediction confidence for each span. The automatic error annotation with XCOMET is based on an LLM that required a larger GPU than our available resources for execution. We run the XCOMET evaluations on CPU where the process is slow, thus we only evaluate the best two models based on the automatic metrics, Llama-3 and Tower.

We show the number of errors ↓ in Table 3 and the MQM scores ↑ in Table 4 for each system. The MQM score summarises the individual errors into a weighted score based on severity (Equation 1). Table 3 shows the number or errors by severity for each model. The total number of errors for the instruction-tuned Llama-3 is: 1914 (en-es), 2910 (en-de), and 1764 (en-ro). The instruction-tuned Tower is: 745 (en-es), 1059 (en-de), and 1632 (en-ro). Instruction-tuned Tower shows fewer critical errors compared to Llama for the three language pairs (en-es, en-de, and en-ro).

Table 4 presents the MQM scores, which show a reduction in critical, major, and minor errors after the instruction tuning phase. In these results, Tower outperforms Llama.

Automatic Error Annotation Analysis We conducted a preliminary error analysis of the automatic error annotation for en-es to assess the quality of XCOMET to label translation errors. A native Spanish speaker with English proficiency served as the annotator. The limited number of examples analysed from the en-es automatic error annotation is because of the lack of a professional medical translator during the preliminary analysis. We show annotation examples between the baseline and instruction-tuned models for Llama-3 and Tower.

Table 5 presents examples of automatic error annotations generated by XCOMET for Llama, QLoRA Llama, and Tower. For Llama, XCOMET identified a critical error with a confidence score of 0.52. Similarly, a critical error in the instruction-tuned Llama was annotated with a confidence of 0.40. While XCOMET produced incomplete annotations, potentially because of over-generation by Llama, it successfully identified code-switched words, such as "*assistant*".

In Tower, XCOMET annotated a major error, "*reconstitución,*" with a confidence of 0.50. For

[5] Unbabel/XCOMET-XL

Model	en-es↓			en-de↓			en-ro↓		
	MIN	MAJ	CRIT	MIN	MAJ	CRIT	MIN	MAJ	CRIT
Llama-3-8B	145	1277	1240	1693	719	938	95	983	1301
QLoRA Llama-3-8B	359	1105	450	2160	295	455	225	844	695
Tower-7B	592	241	15	1266	50	**25**	253	844	695
QLoRA Tower-7B	583	149	**13**	1007	26	26	503	868	**261**

Table 3: Comparing the baseline and QLoRA fine-tuned LLMs with the number of **errors** with the following categories: minor (MIN), major (MAJ), and critical (CRIT).

Model	en-es↑	en-de↑	en-ro↑
Llama-3-8B	35.98	41.29	27.76
QLoRA Llama-3-8B	58.83	59.45	**45.66**
Tower-7B	82.35	80.70	20.11
QLoRA Tower-7B	**86.63**	**84.69**	36.96

Table 4: Comparing the baseline and QLoRA fine-tuned LLMs with **MQM scores** for the en-es, en-de, and en-ro language pairs.

the instruction-tuned Tower, a minor error, "*reconstitu,*" was annotated with a confidence of 0.42. Notably, "*reconstitución*" is the correct term in the MT output with low prediction confidence. A potential solution involves filtering annotations based on a predefined confidence threshold, keeping only high-confidence predictions.

4.3 Quality Estimation

Quality estimation (QE) models predict a quality score for the MT output without using reference translations. QE evaluation can be useful for cases of low-resource language pairs and practical applications, given the lack of reference translations. We use COMETKiwi (Rei et al., 2023) for QE evaluation [6]. COMETKiwi is based on COMET features to train a QE prediction model. The QE model is trained with an annotated multilingual source and corresponding MT outputs to predict quality based on direct assessment (i.e. ranking) or MQM scores.

Table 6 shows the comparison of QE scores for Llama-3 and Tower. The instruction-tuned Tower shows higher QE scores compared to Llama in all language pairs. The QE scores show a similar order in model quality compared to the output of automatic metrics without the need for reference translations.

[6] Unbabel/wmt22-cometkiwi-da

4.4 Discussion and Limitations

Instruction-tuning improves the overall accuracy of terminology and translation quality (e.g. automatic metrics). Instruction-tuned FLAN-T5 (encoder-decoder) has the highest terminology accuracy, but its improvements in translation quality are lower compared to the LLMs. A possible explanation is the difference in parameter size compared to the LLMs, and pre-trained data available for the LLMs. However, to achieve a more accurate evaluation, it is recommended to perform a manual error annotation with professional medical translators.

Both the baseline and instruction-tuned models generate terms defined by our prompts. However, fine-tuning substantially improves accuracy for FLAN-T5, Tower, and Llama-3. Furthermore, Tower includes terminology translation across diverse domains as a component of its tuning tasks.

Llama-3 presents over-generation, producing an excessive amount of tokens with assistant suggestions. For example, in en-es in the test set, the baseline model generates 29,569 tokens, which is reduced to 25,225 tokens after fine-tuning. Examples of this over-generation in Llama-3 include assistant-specific text alongside the expected machine translation output, such as: "*..{source segment} assistant Here is the corrected translation: {MT target segment}...*". However, the instruction-tuned LlaMA-3 also over-generates: *"..I corrected the translation using the provided glossary.assistant Using the glossary..."*, or it continues repeating the MT output. A possible solution is to use a prompt that constrains the model to produce only the target segment. With our current prompt, both Llama-3 models require extra post-processing to extract the MT and avoid biases on the automatic metrics and automatic error annotation. On the other hand, Tower generates 11,034 tokens compared to 10,906 tokens for the instruction-tuned. The MT tasks tuning on Tower improves translation accuracy and avoids over-generation.

Terms	Source, reference, and MT	Annotation
active substance ->principio activo, system ->sistema, fentanyl ->fentanilo	**src**: Ionsys transdermal system delivers the active substance, fentanyl, through the skin.	**critical error**: assistant Here is the translation: Ionsys sistema transdérmico proporciona
	ref: El sistema de liberación transdérmica Ionsys administra el principio activo, el fentanilo, a través de la piel.	
	Llama: Ionsys sistema transdérmico proporciona la sustancia activa, fentanilo, a través de la piel.**assistant Here is the translation: Ionsys sistema transdérmico proporciona** el principio activo, fentanilo, a través de la piel. I corrected the translation using the provided glossary.assistant Using the glossary "system" ->"sistema", I should have translated it to	
active substance ->principio activo, system ->sistema, fentanyl ->fentanilo	**src**: Ionsys transdermal system delivers the active substance, fentanyl, through the skin.	**critical error**: assistant El sistema transdérmico Ionsys libera el
	ref: El sistema de liberación transdérmica Ionsys administra el principio activo, el fentanilo, a través de la piel.	
	QLoRA Llama: El sistema transdérmico Ionsys libera el principio activo, el fentanilo, a través de la piel. **assistant El sistema transdérmico Ionsys libera el** principio activo, el fentanilo, a través de la piel.	
reconstitution ->reconstitución	**src**: Write the date of reconstitution and expiry on the label (expiry is 1 month after reconstitution)	**major error**: reconstitución
	ref: Escriba la fecha de reconstitución y la de caducidad en la etiqueta (la caducidad es 1 mes después de la reconstitución)	
	Tower: Escriba la fecha de reconstitución y el de caducidad en la etiqueta (el de caducidad es 1 mes después de la **reconstitución**).	

Table 5: Examples of **automatic error annotation** for en-es using XCOMET.

Model	en-es↑	en-de↑	en-ro↑
Llama-3-8B	0.513	0.507	0.484
QLoRA Llama-3-8B	0.657	0.619	0.595
Tower-7B	0.840	0.806	0.647
QLoRA Tower-7B	**0.850**	**0.825**	**0.754**

Table 6: Comparing the baseline and QLoRA fine-tuned LLMs with **QE** for the en-es, en-de, and en-ro language pairs.

However, given common limitations on academic computational resources (one GPU) we use small size LLMs (8B) with quantisation, PEFT for tuning our models, and a small split of the EMEA corpus. A limitation of quantisation is the use of pre-trained models with lower precision models that may hurt overall performance. However, SFT in LLMs can be achieved with significantly less data than training from scratch and other domain adaptation approaches (Zhu et al., 2024). The total size of the EMEA corpus is approximately 1M segments.

Automatic error annotation and QE scores offer a detailed evaluation of our language pairs and domain. However, XCOMET shows inaccuracies in terminology annotation, particularly with low-confidence predictions. Furthermore, to validate the reliability of automatic error annotation within the medical domain, a comprehensive analysis involving professional translators is essential. Additionally, XCOMET requires substantial GPU resources.

We use accuracy to evaluate the terms generated in the MT output. The limitation of accuracy is that context is not taken into account (Corral and Saralegi, 2024), for example, translation quality is lowered with high term accuracy in FLAN-T5. A limitation of building our terminology prompt dataset using only exact matches is the potential to miss terms that are expressed differently depending on the context. Furthermore, the coverage of terms and domains within IATE represents a limitation of terminology databases. For example, in the parallel (en-es) segment: *"Posology for MDS/MPD The recommended dose of imatinib is 400 mg/day for adult patients with MDS/MPD."* and *"Posología para SMD/SMP La dosis recomendada de imatinib para pacientes adultos con SMD/SMP es de 400 mg/día"* from IATE the exact term match is "*dose -> dosis*". However, IATE does not contain "*MDS/MPD -> SMD/SMP*" that means "*Myelodysplastic/Myeloproliferative Neoplasm*"[7]. A possible solution is to combine translation terminology databases with medical ontologies, for example, Medical Subject Headings (MeSH)[8], and the Unified Medical Language System (UMLS)[9] that has multilingual features.

[7] https://www.cancer.gov/types/myeloproliferative/hp/mds-mpd-treatment-pdq

[8] https://www.nlm.nih.gov/mesh/meshhome.html

[9] https://www.nlm.nih.gov/research/umls/index.html

5 Conclusions and Future Work

In this study, we show a comparison between baseline LLMs and QLoRA instruction-tuned models in the medical domain for en-es, en-de, and en-ro. We introduce medical terminology from IATE into an instruction-formatted dataset for controlled generation in LLMs. Instruction-tuned models significantly outperform the baseline across automatic evaluation metrics. Furthermore, these models show improved accuracy in terminology translation compared to the baseline.

In particular, the instruction-tuned Tower model presents superior translation quality according to different evaluation methods (automatic metrics, MQM annotation, and QE). Additionally, Tower requires fewer computational resources and less post-processing compared to LLaMA-3.

A limitation of our current evaluation is the reliance on automatic metrics and the limited quality of automatic error annotation. For future work, we will evaluate the baselines with few-shot instead of zero-shot. We will define different prompts for Llama-3 to avoid over-generation. We will perform an evaluation on a balanced test split in terms of the number and type of present terms with respect to the training data. Finally, we will perform a manual error annotation, as automatic metrics may not test for correct terminology generation on the MT output (Haque et al., 2019; Gaona et al., 2023).

Sustainability Statement For the experiments we use a Tesla T4 GPU (16GB) from Azure with an approximate SFT time of 20 hours per model. Instruction-tuning with PEFT tackles issues for scare computational resources (GPUs) for short training time (e.g. one epoch) and small tuning data (60K segments). Moreover, we performed automatic error annotation on the CPU instead of GPU given our academic computational limitations.

From MachineLearning Impact calculator presented in (Lacoste et al., 2019): West-Europe Azure has a carbon efficiency of 0.57 $kgCO_2eq/kWh$. A cumulative of 100 hours of computation was performed on hardware of type T4 (TDP of 70W). Total emissions are estimated to be 3.99 $kgCO_2eq$ of which 100 percent were directly offset by the cloud provider.

Acknowledgments

This work was supported by the ZID of the University of Vienna with Azure cloud credits.

References

ELRC3.0 Multilingual corpus made out of PDF documents from the European Medicines Agency (EMEA), (February 2020) ELRC-SHARE.

Zakaryia Almahasees, Samah Meqdadi, and Yousef Albudairi. 2021. Evaluation of google translate in rendering english covid-19 texts into arabic. *Journal of Language and Linguistic Studies*, 17(4):2065–2080.

Duarte M. Alves, Nuno M. Guerreiro, João Alves, José Pombal, Ricardo Rei, José G. C. de Souza, Pierre Colombo, and André F. T. Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. *Preprint*, arXiv:2310.13448.

Duarte M. Alves, José P. Pombal, Nuno M. Guerreiro, Pedro H. Martins, Joao 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.

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.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *NeurIPS*, volume 13. MIT Press.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Ander Corral and Xabier Saralegi. 2024. Morphology aware source term masking for terminology-constrained NMT. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1676–1688, St. Julian's, Malta. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Abhimanyu Dubey et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Johannes Eschbach-Dymanus, Frank Essenberger, Bianka Buschbeck, and Miriam Exel. 2024. Exploring the effectiveness of llm domain adaptation for business it machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation*.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. 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, Online. Association for Computational Linguistics.

Miguel Angel Rios Gaona, Raluca-Maria Chereji, Alina Secara, and Dragos Ciobanu. 2023. Quality analysis of multilingual neural machine translation systems and reference test translations for the English-Romanian language pair in the medical domain. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 355–364, Tampere, Finland. European Association for Machine Translation.

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.

Rejwanul Haque, Md Hasanuzzaman, and Andy Way. 2019. Investigating terminology translation in statistical and neural machine translation: A case study on English-to-Hindi and Hindi-to-English. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 437–446, Varna, Bulgaria. INCOMA Ltd.

Iikka Hauhio and Théo Kalevi Max Friberg. 2024. Mitra: Improving terminologically constrained translation quality with backtranslations and flag diacritics. In *Proceedings of the 25th Annual Conference of The European Association for Machine Translation*, Switzerland. European Association for Machine Translation. Annual Conference of The European Association for Machine Translation, EAMT ; Conference date: 24-06-2024 Through 27-06-2024.

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.

Frederick Jelinek. 1998. *Statistical Methods for Speech Recognition (Language, Speech, and Communication)*. The MIT Press.

Alexandru-Iulius Jerpelea, Alina Radoi, and Sergiu Nisioi. 2025. Dialectal and low resource machine translation for Aromanian. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7209–7228, Abu Dhabi, UAE. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista Tradumàtica: tecnologies de la traducció*, (12):455–463. Publisher: Universitat Autònoma de Barcelona.

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`.

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.

OpenAI et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jianhui Pang, Fanghua Ye, Longyue Wang, Dian Yu, Derek F. Wong, Shuming Shi, and Zhaopeng Tu. 2024. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Preprint*, arXiv:2401.08350.

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.

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.

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.

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.

Danielle Saunders, Felix Stahlberg, Adrià de Gispert, and Bill Byrne. 2019. Domain adaptive inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 222–228, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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. *Preprint*, arXiv:1910.03771.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

Jiawei Zheng, Hanghai Hong, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *Preprint*, arXiv:2402.15061.

Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. 2024. Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 388–409, Miami, Florida, USA. Association for Computational Linguistics.

A Hyper-parameters

The hyper-parameter values tables for FLAN-T5, Llama-3-8B, and Tower-7B are as follows:

Hyper-parameter	Value
r	8
α	32
Dropout	0.1
Target modules	q, v
Max source length	512
Max target length	512
Batch size	6
Learning rate	$2e-4$
Warm-up steps	0.03
Scheduler type	linear

Table 7: FLAN-T5 seq2seq hyper-parameter values. The upper section contains the QLoRA hyper-parameters, and the lower section contains the overall fine-tuning.

Hyper-parameter	Value
r	64
α	128
Dropout	0.05
Target modules	q_proj, v_proj
Max sequence length	512
Batch size	2
Gradient accumulation	4
Learning rate	$2e-4$
Warm-up steps	0.03
Scheduler type	cosine

Table 8: Llama-3-8B hyper-parameter values. The upper section contains the QLoRA hyper-parameters, and the lower section contains the overall fine-tuning.

Hyper-parameter	Value
r	64
α	16
Dropout	0.1
Target modules	q_proj, k_proj, v_proj, o_proj
Max sequence length	512
Batch size	2
Gradient accumulation	2
Learning rate	$2e-5$
Warm-up steps	0.03
Scheduler type	cosine

Table 9: Tower-7B hyper-parameter values. The upper section contains the QLoRA hyper-parameters, and the lower section contains the overall fine-tuning.

B Instruction Templates

Instruction templates for FLAN-T5, Llama-3 and Tower. The source_term is the source entry from IATE, the target_term is the target entry from IATE, source_language is the source language (i.e. English), target_id is the target language (i.e. Spanish, German, and Romanian), and glossary_type is Glossary with one candidate term pair or Glossaries with several candidate terms.

FLAN-T5 instruction template for a segment with an identified pair of candidate terms. The

prompt is the input for the encoder and the target segment is the input for the decoder:

```
{glossary_type}:
"{source_term}" -> "{target_term}"
...
Translate the source text from {
    source_id} to {target_id} following
    the provided translation glossaries.
{source_id}: {source_segment}
```

FLAN-T5 instruction template with a segment without candidate terms. The prompt is the input for the encoder, and the target segment is the input for the decoder:

```
Translate the source text from {
    source_id} to {target_id}.
{source_id}: {source_segment}
```

Llama-3-8B instruction template for a segment with candidate term pairs:

```
<|begin_of_text|><|start_header_id|>
    system<|end_header_id|>
You are a helpful translation assistant
    .<|eot_id|><|start_header_id|>user<|
    end_header_id|>
{glossary_type}:
"{source_term}" -> "{target_term}"
...
Translate the source text from {
    source_id} to {target_id} following
    the provided translation glossaries.
{source_id}: {source_segment}
{target_id}:<|eot_id|>
<|start_header_id|>assistant<|
    end_header_id|>
{target_segment}<|eot_id|>
```

Llama-3-8B instruction template for a segment without candidate term pairs:

```
<|begin_of_text|><|start_header_id|>
    system<|end_header_id|>
You are a helpful translation assistant
    .<|eot_id|><|start_header_id|>user<|
    end_header_id|>
Translate the source text from {
    source_id} to {target_id}.
{source_id}: {source_segment}
{target_id}:<|eot_id|>
<|start_header_id|>assistant<|
    end_header_id|>
{target_segment}<|eot_id|>
```

Tower-7B instruction template for a segment with candidate term pairs:

```
<|im_start|>user
{glossary_type}:
"{source_term}" -> "{target_term}"
...
Translate the source text from {
    source_id} to {target_id} following
    the provided translation glossaries.
{source_id}: {source_segment}
{target_id}:<|im_end|>
<|im_start|>assistant
{target_segment}<|im end|>
```

Tower-7B instruction template for a segment without candidate term pairs:

```
<|im_start|>user
Translate the source text from {
    source_id} to {target_id}.
{source_id}: {source_segment}
{target_id}:<|im_end|>
<|im_start|>assistant
{target_segment}<|im end|>
```

Lingonberry Giraffe: Lexically-Sound Beam Search for Explainable Translation of Compound Words

Théo Salmenkivi-Friberg[†] and **Iikka Hauhio**[†‡]
[†] Kielikone Oy, Helsinki, Finland
[‡] Department of Computer Science, University of Helsinki, Finland
{theo.salmenkivi-friberg,iikka.hauhio}@kielikone.fi

Abstract

We present a hybrid rule-based and neural method for translating Finnish compound nouns into English. We use a lightweight set of rules to split a Finnish word into its constituent parts and determine the possible translations of those words using a dictionary. We then use an NMT model to rank these alternatives to determine the final output. Since the number of translations that takes into account different spellings, inflections, and word separators can be very large, we use beam search for the ranking when the number of translations is over a threshold. We find that our method is an improvement over using the same NMT model for end-to-end translation in both automatic and human evaluation. We conclude that our method retains the good qualities of rule-based translation such as explainability and controllability while keeping the rules lightweight.

1 Introduction

In this paper, we present a system for translating Finnish compound nouns into English by using a hybrid rule-based and neural method that constructs a trie of possible translation alternatives in a rule-based manner and then selects the translation using beam search.

Unlike in mainstream machine translation research which focuses on sentences and longer texts, our system is instead intended to be used as a fall-back for a dictionary search, providing translations for individual words in case they are not found in the dictionary. This feature is important for languages such as Finnish, Swedish, and German that allow novel *ad hoc* compound words to be formed freely. These compounds are often long and difficult to parse, especially for non-native speakers, rendering traditional dictionary searches unusable for them.

Unfortunately, translating single Finnish compound words into English using current state-of-the-art neural machine translation (NMT) can be quite vexing: at time of writing, the Finnish word ”puolukka-kinuskirahka” (lingonberry caramel quark) gets translated into English as ”lingonberry custard” and ”lingonberry quinusgiraffe” by two popular commercial machine translation systems. These are mistakes no human translator would make, as the Finnish word is unambiguously the compound of ”puolukka” (lingonberry), ”kinuski” (for which caramel is a reasonable translation, even though others can be argued for) and ”rahka” (quark).

“Puolukkakinuskirahka” is a real-world example and a good demonstration of issues neural systems struggle with (cf. Ismayilzada et al., 2024). In principle, translating Finnish compounds to English is an easy task. Both languages share similar rules for simple compounds (see Figure 1) and the constituent parts of the compounds are often common words found in a dictionary or easily translated by an NMT model. However, rule-based translators also struggle with the task, since the compound parts usually have many translations in the target language, and the probability of a wrong translation being chosen grows exponentially as more parts are added to the compound (Forcada et al., 2011; Khanna et al., 2021).

We solve this issue of lexical selection by choosing the best translation given by the rule-based system by scoring the alternatives with an NMT model. Since the number of alternatives can be very large (even hundreds of thousands in some cases), we implement a beam search for searching the space (cf. Cao et al., 2021). We argue that our method combines the good qualities of rule-based translators such as explainability (each word in the translation can be linked back to a dictionary entry) with the versatility of neural methods, not requiring complex rulesets or algorithms for disambiguation.

Section 2 lists prior work with this problem. Section 3 describes our novel methods. Section 4 describes how we evaluated our system on four datasets: a list of food item names, two small forestry-related term banks, and a subset of IATE. Finally, section 5 discusses the next steps and the limitations of our work.

2 Background

In this section, we describe analysis of Finnish and English compound words and cover relevant prior work both in the fields of rule-based and neural machine translation.

Proceedings of Machine Translation Summit XX Volume 1, pages 173–189, June 23-27, 2025

puolukka-kinuskirahka

lingonberry caramel quark

fromage blanc aux airelles et au caramel

Figure 1: An example of compound words in three languages. The structure of the compound is similar for the head-final Finnish and English, while the head-initial French has a reversed structure. Furthermore, French uses gendered and numbered preposition constructions such as “au” and “aux” (a contraction of the plural article and “au”, thus unambiguously referring to the plural “airelles”) and conjunctions such as “et” to convey relationships between components that are left implicit in Finnish compounds. While in this case the use of the dash makes it explicit that “puolukka” and “kinuski” are unrelated and have the same relationship to “rahka”, the added specificity of languages that structure compounds like French cannot be easily deduced from the Finnish or English compound. This makes translating Finnish to English significantly easier than translating it to French.

2.1 Finnish and English Compounds

In this work, we translate expressions that are formed through two mechanisms: compounding or the concatenation of lexemes (producing such constructions as broadsword, single-minded or distance learning) and prefixation or the prepending of a prefix (eg. preschool).

In Finnish, compounds (Finnish: *yhdyssana*) are very common and can be formed productively (Hakulinen et al., 2004, §399). Finnish compound nouns consist of two or more lexemes written together or with a dash, and can be further broken down into so-called “specifier compounds” (*määriteyhdyssana*) consisting of a specifier followed by the head noun, and “sum compounds” (*summayhdyssana*) consisting of two equal parts (Hakulinen et al., 2004, §398). In case of specifier compounds, the specifier can be inflected in any case (Hakulinen et al., 2004, §403).

2.2 Decompounding

As long as the compounds are made of in-vocabulary words, a morphological analyzer such as Voikko (voi, 2025) or Omorfi (Pirinen, 2015) can be used to analyze the compound and return its constituent parts. We call this process “decompounding”. Voikko and Omorfi are both based on a finite-state transducer (FST) (Beesley and Karttunen, 2003) that returns all possible interpretations of the word. The limitation of these tools is that they do not conduct any disambiguation or parsing, i.e., while compound words have a tree-like structure (Hakulinen et al., 2004, §405) (see Figure 2), the tools return a flat list.

The rule-based Finnish–English translator Transmart (Arnola, 1996) always splits compounds in two: the last component and everything else. As Finnish compounds are either symmetric or head-final (Hakulinen et al., 2004, §398), the last part of the compound is often the most important. However, this choice makes translating compounds that have more than two parts impossible unless both parts returned by the analyzer are present in the dictionary.

There are neural alternatives such as Trankit (Nguyen et al., 2021), but presently its Finnish pipeline also returns a flat list while being significantly worse than the FST-based tools.[1] It is also possible to instruct large language models to perform morphological analysis, although the performance on Finnish is still poor (Moisio et al., 2024; Ismayilzada et al., 2024).

2.3 Rule-Based Translation of Compounds

Productively translating Finnish compounds is not a new endeavor, with early attempts such as Transmart (Arnola, 1996) dating back to the 1990s. This was limited by a disambiguation problem that has later been termed lexical selection by Forcada et al. (2011) and Khanna et al. (2021), i.e. selection of the target lexemes corresponding to the source lexemes. While Forcada et al. and Khanna et al. focus on multi-word expressions (MWEs) instead of compound nouns, we argue that their work is still mostly relevant to this work. Khanna et al. present both a data-driven and a rule-based mechanism for lexical selection. The data-driven approach is based on a maximum-entropy model used to generate weighted lexical selection rules. The rule-based approach is a separate dictionary of MWEs that should be translated as a unit, giving the example of ”little brother” and ”big brother” having separate single-word translations in Kyrgyz.

2.4 Constrained Decoding

Constrained decoding is an umbrella term for methods where a generative model such as a large language model is used not to simply generate the most probable text (or in our case, translation) from the set of all possible texts. Instead, at each generation step only a subset of all possible tokens to continue the text with is considered. Constrained decoding has been successfully used for many NLP tasks (Geng et al., 2023). In the field of machine translation, constrained beam search has been used for incorporating glossaries into an NMT system (Hokamp and Liu, 2017; Post and Vilar, 2018; Hu et al., 2019; Hauhio and Friberg, 2024)

In constrained decoding, the constraints are defined using a grammar. If simple enough, this grammar can be represented as a trie (Cao et al., 2021). In more complex situations, a regular expression resulting in a finite-state machine (Hauhio and Friberg, 2024) or a context-free grammar (Geng et al., 2023) can be used. The grammar is used to determine which tokens can be valid continuations for a sequence, and the probabilities

[1]For example, the Trankit Finnish-TDT pipeline analyses “apumekaanikkoaliupseeri” (assistant mechanic NCO) as “apume#kaanikko#aliupseeri” instead of the correct “apu#mekaanikko#ali#upseeri”. Compare with Figure 2.

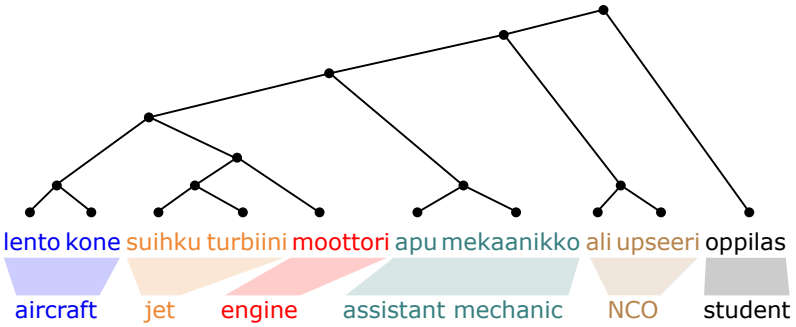


Figure 2: To illustrate the hierarchical nature of Finnish compound words, we use "lentokonesuihkuturbiinimoottoriapumekaanikkoaliupseerioppilas" (aircraft jet engine assistant mechanic NCO student) as an example word. This word is often given in the "longest Finnish words" lists, although it is unclear if it ever has been used in real life (Vartiainen, 2018). While this word is artificially long and would not occur in fluent Finnish text, it is a good example of the recursive nature of Finnish words (Hakulinen et al., 2004, §405). The colored words are found in a dictionary while "konesuihku" (machine jet), "moottoriapu" (motor help), or "mekaanikkoali" (agrammatical) are not. Therefore, merely parsing the word into a flat list of parts is not enough – the tree structure of the word must be maintained such that segments from the dictionary are subtrees of the original tree.

of invalid tokens are set to zero during the decoding. In cases where it is enough for the constraints to appear somewhere in the output such as glossary translation, the grammar contains a wildcard allowing any token to appear. In this work, we use a trie for defining the constraints, and it contains no wildcards, which means that the sequences contained in the trie are the only allowed sequences.

3 Our Method

We present a pipeline for translating Finnish compounds with a hybrid rule-based and neural approach. Our pipeline has the following components:

1. **Morphological Analysis (source language dependent).** This step splits the given word into its component lemmas, returning all possible interpretations of what those lemmas could be. This step results in a list of different ways to split the compound into atomic parts.

2. **Hierarchical Disambiguation (source language dependent).** We decide which atomic compound parts can be combined together into known larger subcompounds that exist in the dictionaries. The result of this step is a list of source components as per Figure 2 (the colored components).

3. **Candidate Generation (target language dependent).** We use a set of rules to generate different spellings, inflections, and other variations of the translations of the compound parts. This accounts for instance for closed versus open compound spelling in the target language.

4. **Token Trie Formation (language independent).** The number of translation alternatives for a compound is at least the product of the number of its components' translation alternatives. This can be in the thousands or tens of thousands, especially when accounting for all the spelling variations we wish to try. To address this, we build a lazily tokenizing trie structure to only construct the translation alternatives that the beam search in the next step will visit.

5. **Beam Search (language independent).** Finally, we search the lazy trie using a beam search to determine a good translation.

3.1 Morphological Analysis

We use the Voikko tool (via the `pyvoikko` Python package) to split the compound words into their constituent parts. As explained in section 2.2, Voikko returns a flat list of the most atomic parts it can find. Knowing the surface forms and order of the parts, we can deduce the corresponding ranges in the string to be analyzed. Some of the ranges may overlap: "maastopaloja" (terrain fires) can be analyzed as *maasto—palo* (terrain fire), *maasto—pala* (terrain piece) or *maa—stop—ala* (earth stop area). For "maastopaloja" we would get the range [1, 3] matching the entry for *maa*, the range [1, 6] matching the entry for *maasto*, the range [4, 7] matching the entry for *stop*, the range [7, 12] twice with one instance matching *pala* and the other *palo* and finally the range [8, 12] matching *ala*.

3.2 Hierarchical Disambiguation

After the analysis, we determine the dictionary translations for the constituent parts. However, in many cases, the dictionary does not only contain the translations for the atomic parts, but also for subcompounds. For example, consider the word "aliupseerioppilas", made of the parts *ali* (sub), *upseeri* (officer), and *oppilas* (student). A word-by-word translation would be "subofficer student". However, the correct translation of *aliupseeri* in English is "non-commissioned officer" or "NCO". Therefore, before determining the translations of the parts, we need

to combine the parts into larger subcompounds in order to find the best translations from the dictionary. This process is further complicated by ambiguity. As "aliupseerioppilas" is made of the subcompound *aliupseeri* and the noun *oppilas*, it should be parsed as *aliupseeri—oppilas*. Interpreting it as *ali—upseerioppilas* is incorrect, even though in our case "upseerioppilas" (officer student) is also present in our dictionary.

Our main solution for the ambiguity is to use the NMT model for choosing the translation from the alternatives using beam search as described in the following sections. However, to reduce the use of resources and ensure useful time-performance, we want the search space to be as small as possible. For this reason, we prune the alternatives using the algorithm described below. We support a case where we have two dictionaries with differing priorities, in our case, a domain-specific glossary and a general dictionary.

1. *Dictionary Query (language independent).* This step looks up translations for the atomic parts and all possible combinations of them in the domain-specific glossary and the general dictionary. For single-letter parts such as "A" in "A-rappu" (A wing), no dictionary query is performed: the single letter itself is returned as the only translation.

   For example, for "aliupseerioppilas", the atomic parts *ali*, *upseeri*, *oppilas*, and the subcompounds *aliupseeri* and *upseerioppilas* are all found in the dictionary. (The full compound *aliupseerioppilas* is not found in the dictionary – if it was, we would return the dictionary translation and not use our system at all.)

2. *Scoring (language dependent).* This step scores translations depending on whether the source term is interpreted as being inflected or not. If it is inflected, it gets a 2.5 point penalty and if its in lemma form, it gets a 1 point penalty. For a given term, only translations with the lowest penalty are kept. In the case of "aliupseerioppilas", all parts are in their base form and receive the penalty of 1.

   We penalize all parts, since we prefer the solutions with fewer compound boundaries and thus longer subcompounds. This is a desirable quality, because compounds can have a different meaning than the sum of their components. Our penalty scheme also prefers doing two extra splits to interpreting a subcompound as inflected, because we find inflected subcompounds to be less likely than a subcompound with a short word that happens to look like a case ending at the end.

3. *Disambiguation (language independent).* This step has two goals: enforcing the domain-specific glossary and finding the best decompounding. We do the first by splitting the compound into dictionary words from the domain-specific glossary and general dictionary so as to maximize the amount of characters covered by words from the domain-specific glossary. We then forget the specifics of what dictionary entry we assigned each range of the compound and consolidate our assigned ranges based on which dictionary they came from. Within each consolidated range, we run another search that finds the decompoundings that have the lowest penalty per step two.

   The word "aliupseerioppilas" does not have domain-specific glossary words, so we treat it as a whole as one consolidated range. We then count the penalties of the possible splits by summing the penalties determined in the previous step: *ali—upseeri—oppilas* has the penalty of 3, while *ali—upseerioppilas* and *aliupseeri—oppilas* both have a penalty of 2. Of these, we return the decompoundings that received the lowest penalty.

   Our analysis would change if we had domain glossary terms. If the user added, for example, "aliupseeri" to the domain-specific glossary, we would have two consolidated ranges: "aliupseeri" and "oppilas". We would then perform this step separately for both of these, but using the glossary for the first range and the general dictionary for the second range. This would result in the single interpretation *aliupseeri—oppilas*. The domain-specific glossary can thus be used for both correcting domain-specific terms and errors caused by incorrect disambiguation.

See Appendix A.1 in general and Algorithms 1 and 2 in particular for a detailed description of this algorithm and Figure 2 for an example of the result of this step.

No edit	"Afro"
Hyphen	"Afro-"
Space	"Afro "
Genitive	"Afro's "
Lower case	"afro"
Lower case + hyphen	"afro-"
Lower case + space	"afro "
Lower case + genitive	"afro's "

Table 1: Different spellings of the word *Afro* generated during the candidate generation.

3.3 Candidate generation

After determining the sequence of compound parts as described above, we generate a list of translation candidates for each part. This is non-trivial, as we have to account for at least capitalization, conveying information from Finnish morphology, and different separators used in different types of English compounds, such as solid, closed, genitive, and hyphenated.

The translation candidates are generated by augmenting the list of dictionary translations determined in the previous phase by including different capitalizations and compound separators: no separator, hyphen ("-"), space

(“ ”), and genitive (“’s ”). For example, the alternatives generated for the word “Afro” are listed in Table 1. The correct spelling and separator depends on the word: e.g., *Afrofuturism*, *Afro haircut*, and *Afro-Asiatic* are all spelled differently. In addition, if the Finnish component was pluralized, we generate the English plural forms of all variations in addition to the singular forms.

The number of alternatives generated can be very high. For example, the Finnish word *afro* has only one translation in our dictionary (“Afro” in different senses), and thus results in the eight alternatives listed in Table 1. However, the word *ali* has five translations (“sub”, “under”, “low-level”, “deputy”, and “hypo”) which result in 40 different alternatives. The word *tulo* has 16 dictionary translations, resulting in 128 alternatives. If the number of alternatives of all the parts in the compound are multiplied with each other, even a word with as little as three compound parts can result in tens of thousands of different possible translations.

To combat this, we prune the number of alternatives by utilizing a list of English prefixes. For words that are not at the end of the compound, we generate candidates as so:

1. We always offer the spelling with a space.
2. We offer the closed compound spelling if the word is on our list of prefixes.
3. We offer the spelling with a hyphen if the word ends with a hyphen.
4. If the Finnish word is plural, we offer the English plural forms. If the Finnish word is in genitive, we offer the English genitive form in addition to the nominative form.

If the word is at the end of the compound, we do not offer the spelling with a space and only offer the spelling with a hyphen if the word itself ends in a hyphen: in that case we assume that the compound ends with some sort of code.

3.4 Decoding

We have two main approaches to decoding. If the number of candidates is smaller than a threshold (less than 400 in our experiments), we rank all of them using the language model in one batch. If we determine that we have too many candidates for this to be efficient, we construct a lazy token trie structure over our candidate translations. Then we implement beam search over the set of all candidates using the trie. Section 3.4.1 describes the token trie structure and section 3.4.2 describes the beam search.

3.4.1 Token Trie

We form a trie of tokens containing all possible tokenizations of the translation alternatives. For example, consider the word “horseshoe”. The optimal SentencePiece tokenization[2] of this word is “_horses hoe”. There is also a multitude of suboptimal tokenizations such as “_horse s hoe”. Our trie includes all possible tokenizations. The rationale for this is that if “horseshoe” is a common word in the model’s training set, the model has likely learned it in the form “_horses hoe”. However, if this was a very rare word, it might be possible that the model has learned its parts “horse” and “shoe”, but not the word as a whole, in which case it makes sense to force a token boundary between the compound parts (“_horse s hoe”). By including the different tokenizations, we allow the NMT model to pick the one it is the most familiar with.

Since the number of combinations is often too large to tokenize at translation time, we perform the tokenization lazily. We form a token trie by first forming a word trie, then a character trie, and finally a token trie. All of these tries are lazy, which means that we only perform tokenization for the branches that are actually searched by the beam search. The details of this step are presented in Appendix A.2.

3.4.2 Beam Search

After forming the trie, we use constrained beam search to search for high-scoring sentences (cf. Cao et al., 2021). In our experiments below, we used a beam size of 50. The beam search is otherwise similar to a regular, unconstrained beam search, but in each step, we remove the hypotheses containing tokens not present in the trie. This limits the output sequences to the sequences encoded in the trie. We also deduplicate the hypotheses such that only the most probable tokenizations per the NMT are retained. The detailed beam search algorithm is presented in Appendix A.3.

4 Evaluation

Our main research question was: “Is this system an improvement over using regular NMT translation for single-word expressions, either in translation quality or in time performance?” To measure the translation quality, we conducted a human evaluation of several compound word datasets translated by both our system and an NMT model decoded with a normal beam search. Along with our human evaluation, we also report automated metrics that we find relevant. Notably, we do not report BLEU as our translation pairs are too short for it to be representative. We also measured the translation time for each of the translated expressions to estimate whether the system is practical.

4.1 Evaluated Systems

We evaluated two systems: our pipeline described in this paper (called “our system” or the “compound translator”), and an NMT model baseline. We ran our pipeline using a commercial dictionary of about 600

[2]Using the English tokenizer of the opus-mt-tc-big-fi-en model.

	Fineli	Hydrology	Forest Soil	IATE	Total
Dictionary hit	118	95	146	1694	2053
No compound translation	19	38	36	114	207
Retained	63	45	20	192	320
Total	200	178	202	2000	2580

Table 2: Terms excluded from evaluation per dataset. Dictionary hit means that the term was found as such in the dictionary. No compound translation means that the term had compound parts that were not found in the dictionary.

000 unique headwords in the Finnish–English direction as the source of compound part translations. This dictionary is designed for human consumption, and no significant changes were made to accommodate the machine translation use case. We scored all alternatives if there were less than 400 and used the beam search otherwise, with beam size set to 50.

We used the Opus-MT Tatoeba Challenge model for Finnish–English[3] (Tiedemann, 2020) as a baseline and also as the scoring model used by our algorithm. For the baseline, we used a beam search of width 50 to match our constrained decoding beam size. We do not compare against other baselines for two reasons: First, the results would not be comparable if the scoring model was different from the baseline model. Second, the chosen model is the state-of-the-art model according to the OPUS-MT Dashboard[4] (Tiedemann and de Gibert, 2023) and choosing a worse model would likely not have given useful information. Furthermore, while we considered using large language models, based on existing literature, they perform poorly in Finnish morphological analysis (Ismayilzada et al., 2024; Moisio et al., 2024) and, while they have demonstrated good translation quality in recent studies (Luukkonen et al., 2024), their inclusion would have required significant computational resources, so ultimately we decided against including them.

4.2 Data

We used the following four test sets for the evaluation:

1. **Fineli Food Composition Database.** We sampled 200 rows from the Basic Package 1 of the national Food Composition Database[5] by the Finnish Institute for Health and Welfare. This dataset includes names of food items. Preprocessing done for this dataset is described in Appendix B.

2. **Forest Hydrology Glossary.** This is a glossary provided by the Finnish Forest Centre. We received an incomplete version of the glossary during its development (Metsäkeskus, 2023a).

3. **Forest Soil Glossary.** As the previous dataset, this was provided by the Finnish Forest Centre in an incomplete state during its development (Metsäkeskus, 2023b).

4. **IATE** We sampled 2,000 terms from the Finnish–English IATE[6]. We used the following parameters when downloading the dataset: all domains, term type term, all reliability levels, and evaluation admitted, preferred, proposed or not specified.

We filtered out terms from the human evaluation in two cases that had to do with the outcome of the dictionary queries. The first case was that in which our system could not parse the term into components found in the dictionary. In real-world use, an NMT translation would be presented to the user. In our experiment, we simply removed these terms.

The much more common case on all data sets was that the compound term was simply found in the dictionary. Our new method reduces to a lemmatizing dictionary search here, both in terms of computational cost and outcome. As such, keeping these terms would be evaluating the contents of the dictionary and not the quality of the system. Again, a real-world user would be presented the lexicographically validated dictionary item that has even stronger accuracy guarantees and upon which a manual intervention is easier. Whereas the first case is quite clearly a undesirable outcome, we see the second case as a not only desirable but even better than coming up with a machine translation at all.

At a deeper level, the rationale for both cases being excluded from our evaluation is that we want to focus on evaluating the compound translator and not the dictionary or the baseline NMT. See Table 2 for the number of terms dropped for the above reasons. Altogether we kept 320 terms.[7]

4.3 Methods

The terms in the datasets were translated with both the proposed system and the Opus-MT model. To measure the improvement in translation quality, we performed both automatic and human evaluation. To measure time performance, we timed the end-to-end time it took for both systems to produce the final translations.

[3] https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-fi-en

[4] https://opus.nlpl.eu/dashboard/index.php?model=top&test=tatoeba-test-v2021-08-07&scoreslang=fin-eng&src=fin&trg=eng (accessed 2025-01-22)

[5] https://fineli.fi/fineli/en/avoin-data (accessed 2025-01-16)

[6] https://iate.europa.eu/home (accessed 2025-01-09); Download IATE, European Union, 2025

[7] Find the evaluated terms at https://github.com/kielikone/mitra-eval-results

4.3.1 Automatic Evaluation

We use the chrF2 (using the `sacrebleu` Python library) and COMET (using the `unbabel-comet` library) metrics. In the case of the Forest Soil data, all our pairs where three words at most, resulting in a BLEU score of 0. As most of our translations are only composed of two or three individual words, we decided against using the word-based BLEU even for datasets where it was non-zero: it would look correct, but not represent the vast majority of the translation pairs.

We used bootstrap resampling to calculate the p-values for the results (using the builtin methods of the `sacrebleu` and `unbabel-comet` libraries).

4.3.2 Human Evaluation

We created an evaluation spreadsheet with rows corresponding to the test words, and columns corresponding to the evaluated systems. We also included a column containing the source language text and another with the reference translation. We randomized the order of the system columns per row. If the NMT model and the proposed system gave the same result, it was presented to the human evaluator only once. If a system gave the reference translation as the result, that translation was not presented to the evaluator and was given the maximal score automatically. If both translations were the same except for differences in capitalization, the one matching in capitalization with the correct translation was presented to the human evaluator and the other was given a grade automatically: if the correctly capitalized got grade 3 (see below), the incorrectly capitalized got grade 2. Otherwise they got the same grade. This deduplication scheme was designed to minimize human labour and systematicise the grading of cases where an obvious grade could be deduced.

We instructed the evaluator to follow the following instructions. The grading scale was devised to allow both comparing the quality of the translations (with the grades having a clear order from worst to best quality) and give useful information regarding the severity of the quality issues relevant for us.

For instance, when translating pikkusuomunokkasärki (a name for the fish species minnow-nase), the following translations would get grades zero to four:

0. *Unsuitable: a translation that has no connection to the original term. 'pinkworm' is an unsuitable translation: it could refer to the name of an animal species, but it has no other commonality with a correct translation.*

1. *Approximate: meaning is conveyed only partially. This category includes translations where the components of the compound have been translated too literally and cases where some part of the compound has been mistranslated. A human that is well-enough versed in the topic can guess what the concept being translated is based on an approximate translation without having knowledge of the source language. 'small scale beak roach' could be an approximate translation: it contains a correct translation for all the components of the Finnish word. A human who knew fish well could perhaps guess at which fish this is: the minnow-nase does seem to have somewhat dense scales, a slightly elongated face and a roach-like body plan generally.*

2. *Spelling mistake: a competent human could produce these words as a translation, but would spell them differently: together / with a space, capitalized / not capitalized, with an accent / without, etc. 'Minnownase' could be an example of a spelling mistake.*

3. *Natural: a competent human could produce this translation. 'minnow-nase' would be a natural translation, as it is the agreed-upon name of this fish.*

Beside the scale, if two system outputs were presented to the evaluator, they were instructed to select one as the better translation if one was better.

Our human evaluator is a linguist with lexicographical expertise employed at our organisation. Evaluation work was carried out during their normal office hours and was compensated according to their normal salary. They are a native Finnish speaker who has an excellent level in English.

We calculated p-values for the results using the Wilcoxon signed rank test.

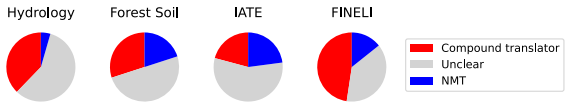


Figure 3: The proportions of terms, for which a system was better than the other according to the human evaluation. This is either inferred from the scores for the translations being different, or directly marked by the evaluator. (See Table 4)

4.4 Results

4.4.1 Automatic Evaluation

The automatic evaluation results are presented in Table 3. With Fineli and Hydrology datasets, the compound translator received significantly higher scores than the NMT baseline, with the chrF2 score increasing from 57.6 to 64.8 (Fineli) and from 47.5 to 52.7 (Hydrology). Similarly, the COMET scores increased from 0.79 to 0.85 and from 0.77 to 0.81, respectively.

On the Forest Soil and IATE datasets, the differences between the two systems were not statistically significant.

4.4.2 Human Evaluation

On the Fineli, Hydrology, and Forest Soil datasets, the compound translator was judged to be better more often

	chrF2			COMET		
	Ours	Opus-MT	p-value	Ours	Opus-MT	p-value
Fineli	**64.8 (64.9 ± 8.2)**	57.6 (57.8 ± 8.6)	0.0060∗	**0.8470**	0.7882	0.0016∗
Hydrology	**52.7 (52.6 ± 9.9)**	47.5 (47.4 ± 9.1)	0.0170∗	**0.8115**	0.7727	0.0090∗
Forest Soil	54.3 (54.6 ± 12.5)	49.9 (50.3 ± 14.9)	0.0949	0.7717	0.7863	0.6104
IATE	50.8 (50.8 ± 4.6)	53.1 (53.2 ± 5.0)	0.0529	0.8012	0.8072	0.3424

Table 3: Automatic evaluation results for the compound translator and the baseline Opus-MT model. For significant results ($p < 0.05$), the better score is bolded.

	Ours	Opus-MT	Unclear
Fineli	**30**	9	24
Hydrology	**17**	2	26
Forest Soil	6	4	10
IATE	40	44	108

Table 4: Number of test words for which the system received a higher grade in human evaluation than the other system. Significant differences are bolded. (See Figure 3)

Dataset	System	Unsuitable	Approximate	Spelling mistake	Natural	p-value
Fineli	Ours	9	15	4	35	0.0000501∗
	Opus-MT	14	13	15	21	
Hydrology	Ours	4	26	0	15	0.00128∗
	Opus-MT	14	18	2	11	
Forest Soil	Ours	3	9	1	7	0.887
	Opus-MT	3	8	2	7	
IATE	Ours	19	107	7	59	0.0584
	Opus-MT	20	100	13	59	

Table 5: The number of each grade the systems received. We used the Wilcoxon signed rank test for significance testing, interpreting our scale as cardinal. We were mainly interesting in whether the difference in grades was statistically significant one way or the other, and so we opted for the two-sided formulation of the test.

	Ours			Opus-MT		
	Avg	Max	Min	Avg	Max	Min
Fineli	0.265	0.715	0.191	**0.157**	**0.417**	**0.111**
Hydrology	0.389	0.622	0.198	**0.138**	**0.402**	**0.0984**
Forest Soil	0.620	0.909	0.410	**0.158**	**0.469**	**0.111**
IATE	0.619	1.06	0.418	**0.137**	**0.393**	**0.0968**

Table 6: Per-term running times (seconds) for compound translation on the different datasets (including the network delays to access the dictionary). Lower time per data set and aggregate function in bold.

than the NMT: 47.6% of terms versus 14.3% (Fineli), 37.8% of terms versus 4.44% (Hydrology); and 30.0% of terms versus 20.0% (Forest Soil). On the larger IATE dataset, the NMT was judged to be better slightly more often than the compound translator: 25.6% of terms versus 20.9%. These results are presented in Figure 3 and Table 4. Like with the automatic evaluation, only the difference in Fineli and Hydrology datasets was statistically significant.

The number of each grade received by the systems is presented in Table 5. We find that the compound translator produced at least as many natural translations as the NMT on every dataset and strictly more on the Fineli and Hydrology terms. On the other end of the scale, it produced no more unsuitable translations than the NMT model on any dataset and again strictly fewer on the Fineli, Hydrology and IATE terms.

We also analyzed how the terms moved from one grade group to another and rendered Sankey diagrams based on this data (see Appendix C). The Hydrology dataset was particularly clean in this, as every translation by the compound translator was rated as at least as good as the corresponding NMT translation (see Figure 8). The case of Fineli (Figure 7), Forest Soil (Figure 9), and IATE (Figure 10) data is more ambiguous, but some patterns do emerge. In particular, terms with spelling mistakes in the NMT translations tend to often flow upward to the highest translation category. In the case of the Hydrology data, spelling mistakes are even completely eliminated. In the IATE dataset, most spelling mistakes from the compound translator were terms that were translated cleanly by the NMT. At least some of these are proper names built up from common nouns (eg. Cohesion Report, Arms Trade Treaty) which have the correct capitalization per the NMT and are spelled fully in lowercase by the compound translator.

4.4.3 Time Performance

The average, maximum, and minimum translation times are presented in Table 6. The compound translator is significantly slower, requiring more than 0.2–0.7 seconds per one test word on average. The Opus-MT NMT model, on the other hand, uses less than 0.2 seconds for each translation on average. There is a significant difference between the Fineli and Hydrology test sets and the Forest Soil and IATE test sets with the latter two requiring over 0.6 seconds on average, while the other two required less than 0.4 seconds on average.

The compound translator makes a request to the dictionary database and performs morphological analysis, both of which take time not required by the NMT model. This added time is an artefact of our test setup and not a constant of our method. Thus, these numbers do not generalize to other setups that, for example, place the database on the same system as the translator and therefore avoid network delay.

5 Discussion

In this section we discuss the implications of our research and take a look at the next steps and further research.

5.1 Scoring is More Practical than Rules

The decoder of a sequence-to-sequence model is a language model and can be used to calculate probabilities of texts. This allows our rule-based component (see Sections 3.2 and 3.3) to be very simple. Instead of focusing on linguistically sound rules that provide good translations, we can just produce as much different alternatives as possible, and filter them with the NMT model. More specifically, we can leave out almost all rules related to disambiguation and focus on rules related to morphological analysis and generation. Since morphological analysis and generation is a much more common task than disambiguation, it is often possible to use preexisting morphological wordlists and libraries (such as Voikko in our case). This radically simplifies the process of writing the ruleset, which is often the bottleneck of development in rule-based machine translation.

Despite our simplified rules, the system maintains, in our opinion, the best qualities of rule-based translators: explainability and controllability. Each translation alternative can be traced back to the dictionary entry that produced it. Furthermore, the user can control the output by editing the dictionary and by deciding which domain-specified glossaries to use.

5.2 The Effect of the Dictionary

The performance of the compound translation is highly dependent on the dictionary. We noticed that the majority of test words in our test datasets were found in the dictionary we used (see Table 2). The compound translator performs better with datasets that have a lower dictionary coverage: the difference between it and the baseline was greatest on the Fineli dataset that had 59% dictionary coverage. Both Forest Soil and IATE had ca. 90% dictionary coverage, but we found no significant difference between the compound translator and the baseline. We argue that while the translator did not improve performance on all datasets, it improved it where it mattered: on the dataset not covered by the dictionary.

While this may be an artifact of the chosen general dictionary and the datasets, we hypothesize that the reason for this might also be that the words in the Forest Soil and IATE datasets are on average more unintuitive and not merely the sum of their parts, which both causes the compound translator to perform poorly, but also is the reason for their inclusion in the dictionary: unintuitive words that cannot be translated productively are more important for the dictionary user. The Fineli dataset, on the other hand, contains names of food items that are typically constructed systematically by listing the same ingredients in all languages.

5.3 Next steps

Currently, our system assumes that the word order of the source and target languages is the same, with a limited number of separators between the compound parts. However, to support languages with differing word order and more complex multi-word expressions (MWE), this assumption does not hold anymore (see Figure 1). This might require replacing the trie structure with a more complex datastructure such as a finite-state machine (FST) (cf. Hauhio and Friberg, 2024) encoding the different word orders.

In this paper, we focused on translating single compound words. However, translating sentences and whole texts is a much more common task in machine translation. We hypothesize that our system could be combined with a terminology-constrained translation algorithm (Hauhio and Friberg, 2024; Bergmanis and Pinnis, 2021; Nieminen, 2024) by scanning the sentence for compounds and adding them as constraints for the translator. In the case of FST-based algorithms, the token trie produced by our system might even be directly incorporated into the finite-state machine.

We hypothesize that many of the issues NMT models have with long compound words are caused by unoptimal tokenization. For example, the word used in the title of this paper, "puolukkakinuskirahka", was translated as "lingonberry giraffe" likely because the model confused the substring "kirah" with the word "kirahvi" *giraffe*. This issue might be mitigated simply by tokenizing the string differently and forcing a token boundary between the compound parts (in this case, between "kinuski" and "rahka").[8] See Section 3.4.1 for more discussion about tokenization.

5.4 Limitations

A major limitation is that we only compared the compound translator to one other system. The choice of the NMT system affects the results considerably. In particular, the model we used has been predominantly trained with full sentences, and its quality may have degraded when applied to single-word source texts. Furthermore, the NMT model was used with a 50-wide beam. It is possible that a greedy search or a smaller beam size could have returned better results (cf. Yang et al., 2018). As noted in Section 5.2, the choice of the dictionary used is also important. In this work, we only used one model and one dictionary, which limits the generalizability of our results.

Another limitation is that our evaluation is relatively narrow: We only used chrF2 and COMET for automatic evaluation, and COMET is intended for scoring full sentences, so it might not be as indicative for single-word translation quality. For manual evaluation, we only had one reviewer. Furthermore, our test datasets were quite small since we had to drop many terms that were already present in our dictionary.

6 Conclusions

In this paper, we presented a system for translating Finnish compound nouns into English using a hybrid rule-based and neural approach. According to our automatic and human evaluation, our system performs better on average than the baseline NMT model on two of our four test sets and has the same performance on the other two test sets. Unlike the NMT model, our system is explainable and controllable, allowing the user to see the dictionary entries the translation is based on, and to fix possible errors by simply modifying the glossary used for translation. While we limited our scope to a single language pair and only compound words, we see promise in our methods to be usable in many kinds of different hybrid rule-based and neural systems.

Carbon Impact Statement

The sum of the translation times of all terms was a hair under 25 minutes, but it does not account for network delays for round trips to the laptop coordinating the translation or partial re-runs that we had to do. In total we estimate that we used no more than an hour of GPU time on an NVIDIA Tesla T4 GPU rated at a 70W maximum power draw to translate a little over 5000 terms. In practice these experiments were running on Amazon g4dn.xlarge instances in the Stockholm region. We used two M3 MacBook Pros and their built-in GPUs and the aforementioned g4dn.xlarge instances for the development of the system and trained no new models. We find it likely that normal development activities on our two laptops, our CI systems and our two g4dn.xlarge staging instances (running mostly idle) over the course of multiple months of development are the larger energetic cost. The model we used is also comparatively small, coming in at 236 megaparameters.

Overall, this results in an approximated less than 60 g CO2e emissions based on the `https://calculator.green-algorithms.org/` online calculator. Of this, most is for the testing during the development of the system and only less than 1 g is for the experiments.

Acknowledgments

Both TS and and IH are funded by Kielikone Oy. IH is additionally funded by the Doctoral Program in Computer Science at the University of Helsinki. We would like to thank Pekka Kauppinen for evaluating the translated compounds, and the peer reviewers for providing valuable feedback and suggestions for further research and possible ablation studies.

[8] We tested this one case and found that "puolukkakinuskirahka" was tokenized as `['_puol', 'ukka', 'kin', 'us', 'kir', 'a', 'hka']`. When we instead tokenized it as `['_puol', 'ukka', 'kin', 'us', 'ki', 'rah', 'ka']`, the NMT model translated it as "lingonberry caramel". While this is not the desired "lingonberry caramel quark", it is better than the translation "lingonberry giraffe" given with the default tokenization. In both of these cases, we used the beam width of 5.

References

2025. Voikko – free linguistic software and data for finnish. https://voikko.puimula.org/. Accessed: 2025-01-13.

Harri Arnola. 1996. Kielikone Finnish-English MT system "TranSmart" in practical use. In *Proceedings of Translating and the Computer 18*.

Kenneth R Beesley and Lauri Karttunen. 2003. Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*, pages 359–375.

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.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25:127–144.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952, Singapore. Association for Computational Linguistics.

Auli Hakulinen, Maria Vilkuna, Riitta Korhonen, Vesa Koivisto, Tarja-Riitta Heinonen, and Irja Alho. 2004. *Iso suomen kielioppi*. Suomalaisen Kirjallisuuden Seura.

Iikka Hauhio and Théo Friberg. 2024. Mitra: Improving terminologically constrained translation quality with backtranslations and flag diacritics. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 100–115, 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.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. 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 839–850.

Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Lonneke van der Plas, and Duygu Ataman. 2024. Evaluating morphological compositional generalization in large language models. *arXiv preprint arXiv:2410.12656*.

Tanmai Khanna, Jonathan N Washington, Francis M Tyers, Sevilay Bayatlı, Daniel G Swanson, Tommi A Pirinen, Irene Tang, and Hector Alos i Font. 2021. Recent advances in apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*, 35(4):475–502.

Risto Luukkonen, Jonathan Burdge, Elaine Zosa, Aarne Talman, Ville Komulainen, Väinö Hatanpää, Peter Sarlin, and Sampo Pyysalo. 2024. Poro 34b and the blessing of multilinguality. *Preprint*, arXiv:2404.01856.

Metsäkeskus. 2023a. Metsähydrologia-sanasto. Received: 2023-08-04.

Metsäkeskus. 2023b. Metsämaa-sanasto. Received: 2023-11-28.

Anssi Moisio, Mathias Creutz, and Mikko Kurimo. 2024. LLMs' morphological analyses of complex FST-generated Finnish words. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 242–254, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Tommi Nieminen. 2024. Adding soft terminology constraints to pre-trained generic MT models by means of continued training. In *Proceedings of the First International Workshop on Knowledge-Enhanced Machine Translation*, pages 21–33, Sheffield, United Kingdom. European Association for Machine Translation (EAMT).

Tommi A Pirinen. 2015. Omorfi—free and open source morphological lexical database for finnish. In *Proceedings of the 20th Nordic conference of computational linguistics (NODALIDA 2015)*, pages 313–315.

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.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation 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 1 (Long Papers)*, pages 1314–1324.

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.

Jörg Tiedemann and Ona de Gibert. 2023. The OPUS-MT dashboard – a toolkit for a systematic evaluation of open machine translation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 315–327, Toronto, Canada. Association for Computational Linguistics.

Vilma Vartiainen. 2018. Pisin sana. *Kielikello*.

Yilin Yang, Liang Huang, and Mingbo Ma. 2018. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3054–3059, Brussels, Belgium. Association for Computational Linguistics.

A Algorithms

A.1 Hierarchical Disambiguation

We call the process that parses our compound into subcomponents found in our dictionaries hierarchical disambiguation, as it respects the hierarchy of the subcompounds (see Figure 2). Besides respecting the hierarchy, we wish the resultant parse to have two other desirable properties: respect our domain glossary and be otherwise linguistically plausible. Our solution to this is composed of two successive dynamic programming algorithms.

The first one or Algorithm 1 concerns itself with enforcing the domain glossary: we consider it less desirable to fail by not using domain glossary terms than parsing the hierarchy suboptimally due to those terms. It is a failure modality that is easily diagnosed by translating without the domain glossary and easily fixed by ammending the domain glossary. Given the ranges in the compound that entries from the domain glossary and entries from the general dictionary represent, Algorithm 1 splits the compound into domain glossary ranges and general dictionary ranges such that the number of characters attributed to domain glossary ranges is maximized. This does not yet tell us what is the best way to parse the subcompounds, only that some solution exists.

The second one or Algorithm 2 is then run individually on each of the subcompounds produced by the previous step. Each run of Algorithm 2 is only given access to terms coming from the dictionary matching that subcompound, thus enforcing the domain glossary whenever some subrange can be parsed with only domain terms. Algorithm 2 produces the parse that minimizes the penalties associated with the ranges.

At the end of this search, we get a sequence of dictionary entries (and the grammatical forms they were inflected in) that match the compound word that was given. Alternatively, we get no sequence at all, indicating that the dictionary we have can not be used to produce a translation for the given compound. This happens early, right after our first dynamic programming search (Algorithm 1). In this situation, we find it wise to fall back onto a normal NMT system, potentially using a constraint-translation method for honoring the domain-specific glossary (cf. Hauhio and Friberg, 2024). If a sequence is produced, it has some desirable qualities. In particular, it represents the interpretation of the compound word that had the most characters covered by domain-specific terms. If there is only one way to choose the ranges to maximally cover from our domain glossary, the parse is further the one considered linguistically the most credible per our scoring. Given this parse, we can proceed onto generating the set of candidate translations, as we see in Section 3.3.

A.2 Token Trie

For the constrained beam search described in Section 3.4.2, we need to know what tokens an incomplete translation can continue with. As we may have exponentially many translation candidates in regard to the number of compound components, we wish to have a data structure that can be built lazily, focusing only on the translation candidates the model seems to prefer. Finally, the data structure needs to somehow map from our compound components to tokens. This is made difficult by SentencePiece, which expects look-ahead to the next piece of punctuation. This can be in the next compound component if we are unlucky. We would need to know the full translation in order to tokenize it. The NMT model is trained on optimal SentencePiece tokenizations, and we do not trust it to rank suboptimal tokenizations reliably.

The solution we present here is a multi-level prefix trie. Using a trie for this kind of constraint decoding is a well-known technique (Hu et al., 2019; Hauhio and Friberg, 2024). Some previous works do incorporate a level of dynamism into the trie itself: Cao et al. (2021) use control tokens to either enable or disable constrained decoding from the trie, effectively amounting to an infinite nested trie structure. However, all previous works we have found either have a small-enough set of options to tokenize them fully at translation time (Hu et al., 2019; Hauhio and Friberg, 2024) or know all the options ahead of time (Cao et al., 2021). Knowing all the items in the trie ahead of time allows Cao et al. (2021) to simply tokenize everything before their inference stage and have tokenization take as much time as it needs. We need to deal with tokenization while we are running

Algorithm 1 The constraint coverage algorithm. Takes as input a sequence of ranges that belong to dictionary entries from a domain-specific glossary, a sequence of ranges that belong to dictionary entries from a general dictionary and the length of the full string these ranges point to. Returns either null if there is no way to select non-overlapping ranges that cover the whole of the string or a selection of ranges that do. If it returns such a selection of ranges, they are each annotated with a boolean value indicating whether it belongs to the domain specific glossary. This selection is guaranteed to have the maximal number of character indices covered by ranges from the domain specific glossary.

```
1:  procedure COVERAGE(glossary_ranges, normal_ranges, length)
2:      coverages ← [0]
3:      solutions ← [([0, 0], False)]                ▷ Here [0, 0] is a range
4:      glossary_lookup ← empty hash map
5:      for all r ← glossary_ranges do
6:          if r.stop ∉ glossary_lookup.keys() then
7:              glossary_lookup[r.stop] ← []
8:          end if
9:          glossary_lookup[r.stop].append(r)
10:     end for
11:     normal_lookup ← empty hash map
12:     for all r ← normal_ranges do
13:         if r.stop ∉ normal_lookup.keys() then
14:             normal_lookup[r.stop] ← []
15:         end if
16:         normal_lookup[r.stop].append(r)
17:     end for
18:     for all i ← [1, text_length] do
19:         best_coverage ← −∞
20:         solution ← null
21:         for all r ← glossary_lookup[i] do
22:                                                  ▷ Note the +|r|
23:             if coverages[r.start] > −∞ ∧ best_coverage ≤ coverages[r.start] + |r| then
24:                 best_coverage ← coverages[r.start] + |r|      ▷ likewise
25:                 solution ← (r, True)
26:             end if
27:         end for
28:         for all r ← normal_lookup[i] do
29:             if coverages[r.start] > −∞ ∧ best_coverage ≤ coverages[r.start] then
30:                 best_coverage ← coverages[r.start]
31:                 solution ← (r, False)
32:             end if
33:         end for
34:         solutions.append(solution)
35:         coverages.append(best_coverage)
36:     end for
37:     if solutions ends with null then
38:         return null                              ▷ No solution was found
39:     end if
40:     selected_ranges ← []
41:     i ← |solutions|-1
42:     while i > 0 do
43:         selected_ranges.append(solutions[i])
44:         i ← solutions[i][0].start
45:     end while
46:     reverse(selected_ranges)
47:                 ▷ Consolidate subsequent ranges with the same source.
48:     return consolidate(selected_ranges)
49: end procedure
```

Algorithm 2 The hierarchical parsing algorithm, made up of two procedures. The main procedure PARSE takes as argument the ranges that are attributed to dictionary entries that are from the correct dictionary and within relevant_range, their matching penalties and the range of the subcompound it is operating in. It returns the set of lists of ranges that the subcompound can be parsed into with the minimal penalty. UNRAVEL backtracks through the dynamic programming data structure and recursively produces the aforementioned set, to be returned by PARSE. It is the caller's responsibility to match the ranges to dictionary entries.

```
1:  procedure PARSE(ranges, penalties, relevant_range)
2:      range_lookup ← empty hash map with default value of []
3:      for all r, p ← zip(ranges, penalties) do
4:          range_lookup[r.stop].append((r, p))
5:      end for
6:      cost_table ← empty hash map with default value of ∞
7:      cost_table[relevant_range.start] ← 0
8:      solutions ← empty hash map with default value of []
9:      for all i ← sorted(range_lookup.keys()) do
10:         for all r, p ← range_lookup[i] do
11:             if cost_table[r.start] + p < cost_table[i] then
12:                 cost_table[i] ← cost_table[r.start] + p
13:                 solutions[i].clear()
14:             end if
15:             if cost_table[i] < ∞ ∧ cost_table[r.start] + p = cost_table[i] then
16:                 solutions[i].append((r, p))
17:             end if
18:         end for
19:     end for
20:     if solutions[relevant_range.stop] = [] then
21:         return ∅
22:     end if
23:     return UNRAVEL(solutions, relevant_range.stop, relevant_range.start)
24: end procedure
25: procedure UNRAVEL(solutions, i, start)
26:     if i = start then
27:         return {[]}
28:     end if
29:     result ← ∅
30:     for rhs ← solutions[i] do
31:         for lhs ← UNRAVEL(solutions, rhs[0].start, start) do
32:             result.add(concatenate(lhs, rhs[0]))
33:         end for
34:     end for
35:     return result
36: end procedure
```

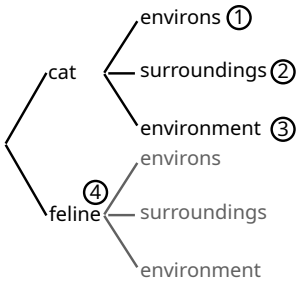


Figure 4: Compound trie. The lazy search has not explored continuations to “feline”. The circled numbers are search states held by the character trie (see Figure 5).

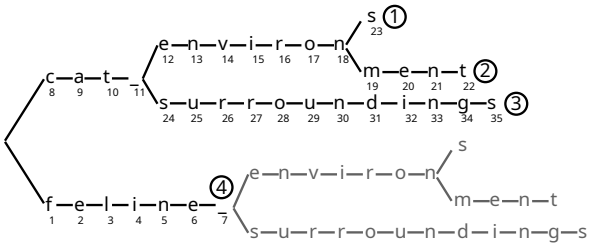


Figure 5: Character trie. The character trie holds references to the compound trie (see Figure 4) in its cells (circled numbers). Note the unexplored continuations of feline.

inference, and so even if our trie is multiple orders of magnitude smaller, it is still a bottleneck.

Our trie has three levels: on the bottom level, we have a lazy search tree in the compound space (see Figure 4). It handles the complexities of generating spelling variations and can generate the set of all follow-up search states with one translation of a compound locked in. It indicates that the search is complete by returning an empty set.

On top of this lazy search tree, we build a character-level trie (see Figure 5). Externally, it has a similar interface: it has a method to list all characters that can continue a string along with indices in the trie that match these continuations. Internally the cells of this trie hold references to the compound search states. If the possible continuations of a cell that holds no references to the compound search are queried, the outgoing edges from that cell are returned. Otherwise, the compound search states are asked to generate their follow-ups and nodes are added to the character trie accordingly. The references that the original cell was holding are cleared and any new cells matching an unexplored search state are pointed to it. The character trie is initialized with a single cell holding a reference to an unexplored compound search tree. Thus we can abstract away the compound formation completely and deal with characters without losing the laziness of the compound search tree.

Finally, on top of the character trie, we build the token trie (see Figure 6). In a similar manner, it exposes a method to list all tokens that can continue a token sequence along with indices into the token trie that match them. Cells in the token trie hold references to the character trie and unravel it just enough to know what tokens can continue a sequence, much like the character trie itself does with the lazy search tree. In practice, the token trie holds a character-level prefix trie of all the tokens in the vocabulary, allowing for fast elimination of tokens that are not allowed given the allowed characters that the character level trie produces. This third level allows us to answer queries about the token-level representation of the translation without fully knowing it in advance: among the options that the token trie generates, the correct SentencePiece tokenization is bound to exist and selecting it is the task the NMT model has been trained on.

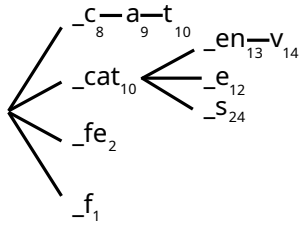


Figure 6: Token trie. The token trie holds references to the character trie (see Figure 5) in its cells (subscript).

A.3 Beam Search

The beam search algorithm is presented as Algorithm 3.

B Data Preprocessing

We sampled 200 rows from the Basic Package 1 of the national Food Composition Database in Finland[9]. It required some non-trivial pre-processing. Each row of the files foodname_FI.csv and foodname_EN.csv in the package contains a description of an ingredient or a food item such as ’FLOUR MIXTURE, FOR BREAD ROLLS, WHEAT FLOUR, WHEAT GROATS, RYE’. We split this string on commas and only kept the first chunk, as it seemed to match better across languages. These string pairs were filtered to only keep the ones where the Finnish side contains only dashes, characters A-Z and Å, Ä and Ö. Selected strings were then lowercased. From the lowercased string pairs, a sample of 200 was randomly drawn. The matching of the Finnish and English terms was somewhat imperfect, as noted by our evaluator.

C Sankey Diagrams

We produced the Sankey diagrams in Figures 7, 8, 9, and 10 to visualize how our system changed the translations when compared to the NMT model. The diagrams have the evaluation result of the NMT model on the left and our system on the right.

[9] https://fineli.fi/fineli/en/avoin-data

Algorithm 3 Constraint beam search algorithm used in this work. Unlike the ones used in (Hauhio and Friberg, 2024), (Hu et al., 2019) and (Post, 2018), it is rooted such that the constraint starts at the start of the string and spans the whole string. Takes as arguments the token trie object, the width of the beam and a function LM that calls the language model on token sequences. Returns a tuple with the tokens of the best hypothesis as the first element and its score as the second. The book-keeping to match the selected dictionary entries to the hypotheses is ommitted for brevity. In practice, the token trie would know the matching hierarchical analysis for any final state. The analysis for a finished translation can then be deduced by iterating through the trie token by token.

```
procedure TRIEBEAMSEARCH(trie, beam_size, LM)
    generations ← [[start token]]
    trie_indices ← [0]
    ongoing_scores ← [0]
    finished_hypothesis ← null
    while |trie_indices| > 0 do
        continuations ← []
        for all hypothesis ← [0, |trie_indices|) do                ▷ Generate candidate continuation for all hypotheses.
            for all new_idx, token ← trie.children(trie_indices[hypothesis]) do    ▷ Continuations coming from trie.
                continuations.append((token, hypothesis, new_idx))
            end for
            if trie.finished_hypothesis_at(trie_indices[hypothesis]) then    ▷ Continuations that finish hypotheses.
                continuations.append((eos token, hypothesis, null))
            end if
        end for
        scores ← LM(generations) + ongoing_scores
        relevant_scores ← []
        for all c ← continuations do
            relevant_scores.append(scores[c[1]][c[0]])
        end for
        score_order ← relevant_scores.argsort()
        reverse(score_order)
        for all idx ← score_order do                    ▷ Update the finished hypothesis.
            if finished_hypothesis ≠ null and finished_hypothesis[1] > score_order[idx] then
                break
            end if
            if continuations[idx][2] = null then                ▷ This must be a finished hypothesis.
                finished_hypothesis ← (generations[continuations[idx][1]] + [continuations[idx][0]], relevant_scores[idx])
            end if
        end for
        new_trie_indices ← []
        new_generations ← []
        new_ongoing_scores ← []
        for all idx ← score_order do
            if |new_generations| = beam_size or finished_hypothesis[1] ≥ score_order[idx] then    ▷ Traditional cut-off
                break
            end if
            if trie_idx ∈ new_trie_indices then    ▷ We have got a cheaper way to tokenize this string.
                continue
            end if
            new_trie_indices.append(continuations[idx][2])
            new_generations.append(generations[continuations[idx][1]] + [continuations[idx][0]])
            new_scores.append(relevant_scores[idx])
        end for
        trie_indices ← new_trie_indices
        generations ← new_generations
        ongoing_scores ← new_ongoing_scores
    end while
    return finished_hypothesis
end procedure
```

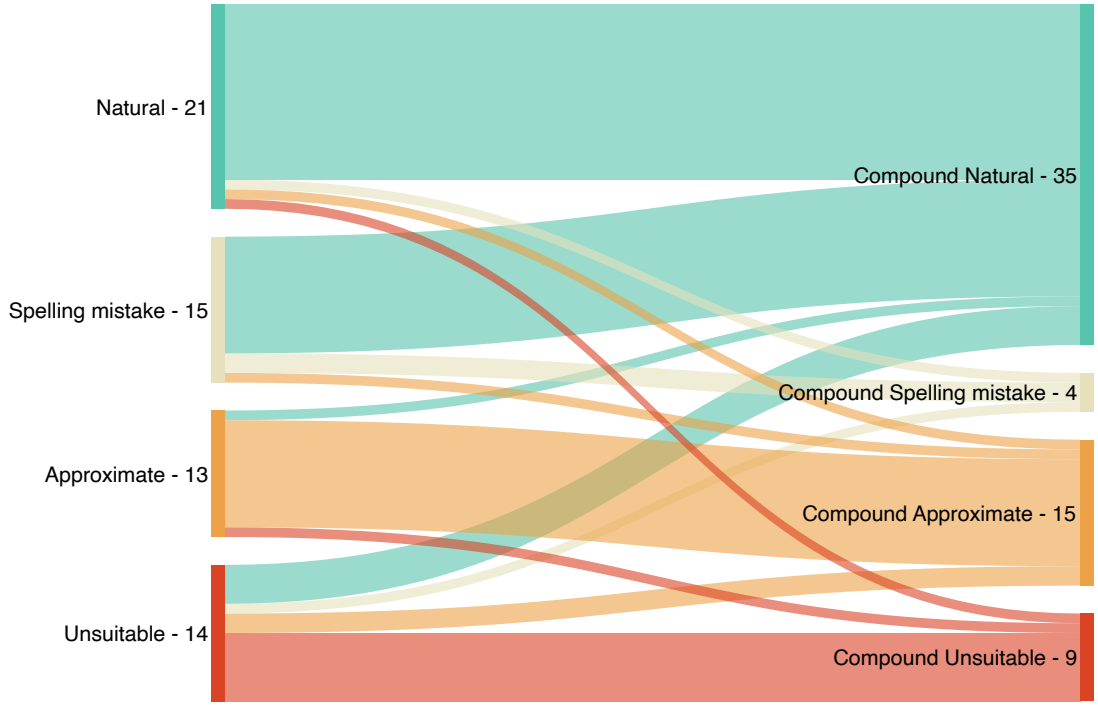


Figure 7: Distribution of the FINELI terms with NMT evaluation on the left and compound translator evaluation on the right.

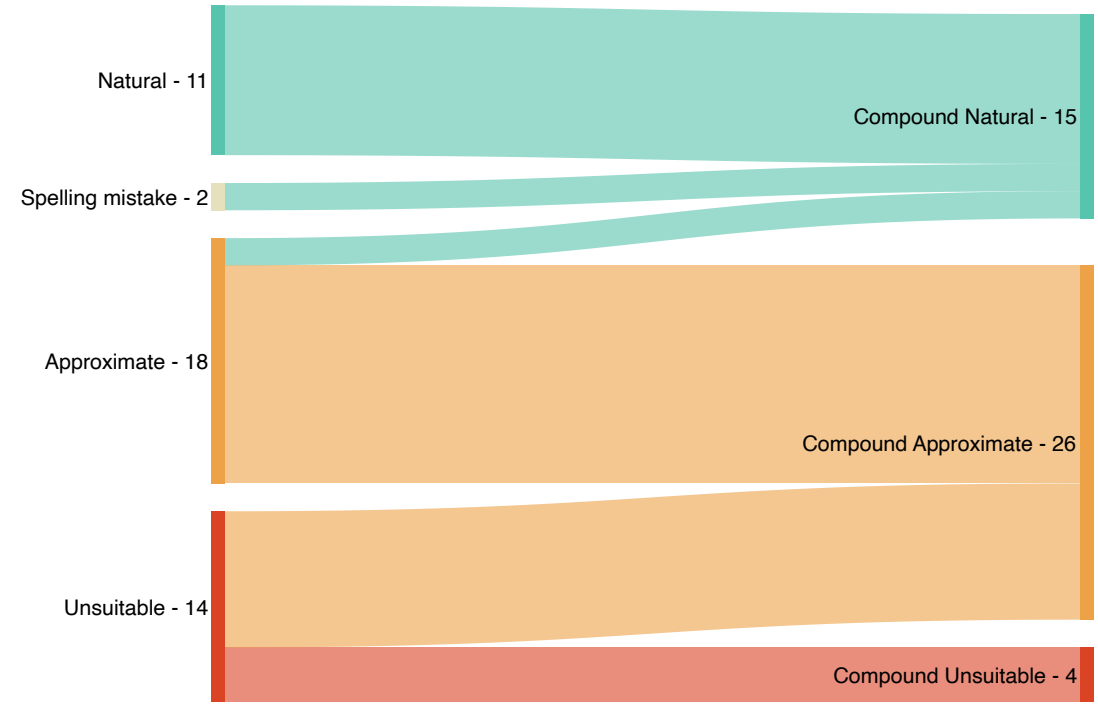


Figure 8: Distribution of the Hydrology terms with NMT evaluation on the left and compound translator evaluation on the right.

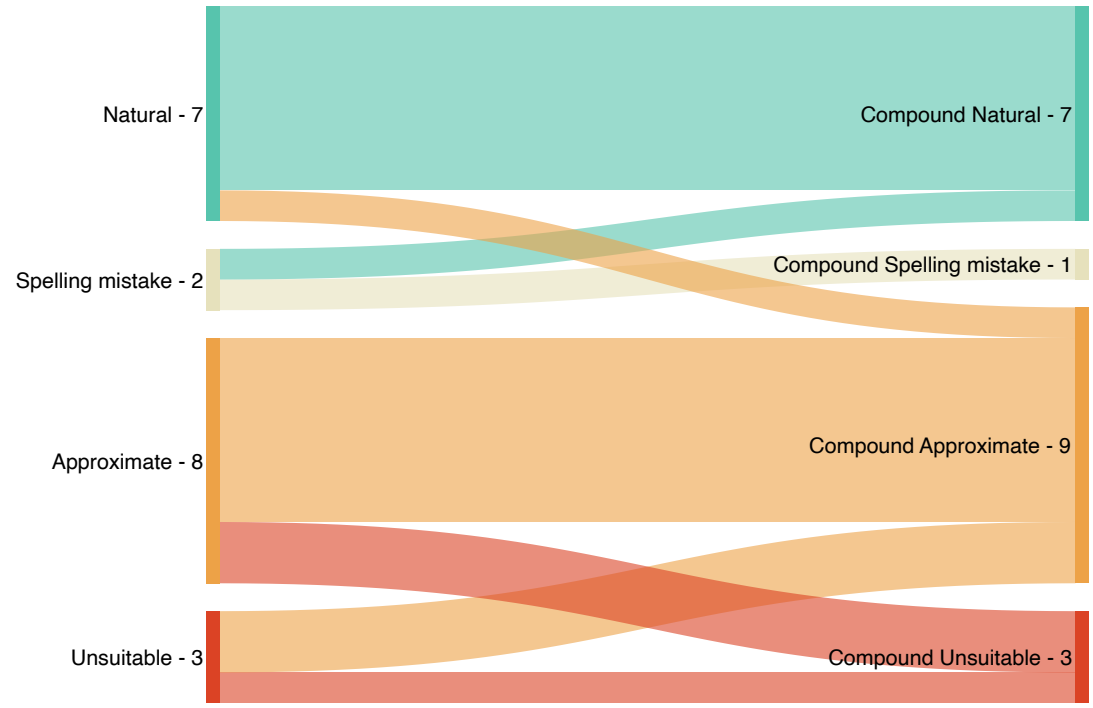


Figure 9: Distribution of the Forest Soil terms with NMT evaluation on the left and compound translator evaluation on the right.

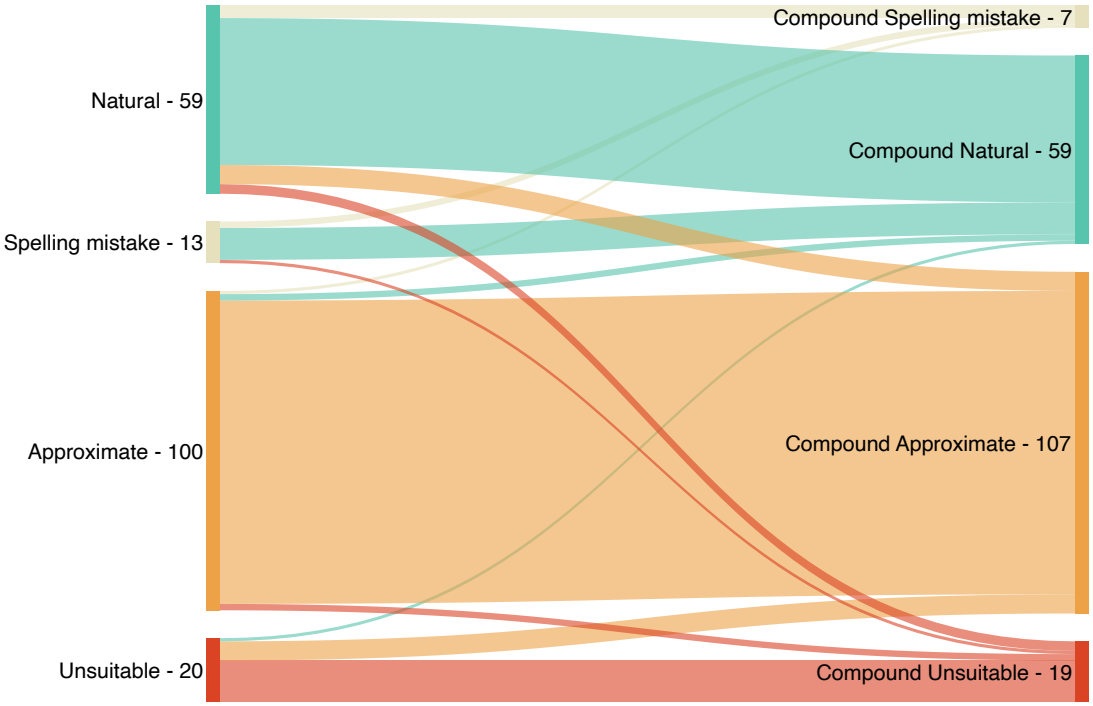


Figure 10: Distribution of the IATE terms with NMT evaluation on the left and compound translator evaluation on the right.

Testing LLMs' Capabilities in Annotating Translations Based on an Error Typology Designed for LSP Translation: First Experiments with ChatGPT

Joachim Minder and **Guillaume Wisniewski** and **Natalie Kübler**
Université Paris Cité, ALTAE, F-75013 Paris, France
Université Paris Cité, CNRS, Laboratoire de linguistique formelle, F-75013 Paris, France

Abstract

This study investigates the capabilities of large language models (LLMs), specifically ChatGPT, in annotating MT outputs based on an error typology. In contrast to previous work focusing mainly on general language, we explore ChatGPT's ability to identify and categorise errors in specialised translations. By testing two different prompts and based on a customised error typology, we compare ChatGPT annotations with human expert evaluations of translations produced by DeepL and ChatGPT itself. The results show that, for translations generated by DeepL, recall and precision are quite high. However, the degree of accuracy in error categorisation depends on the prompt's specific features and its level of detail, ChatGPT performing very well with a detailed prompt. When evaluating its own translations, ChatGPT achieves significantly poorer results, revealing limitations with self-assessment. These results highlight both the potential and the limitations of LLMs for translation evaluation, particularly in specialised domains. Our experiments pave the way for future research on open-source LLMs, which could produce annotations of comparable or even higher quality. In the future, we also aim to test the practical effectiveness of this automated evaluation in the context of translation training, particularly by optimising the process of human evaluation by teachers and by exploring the impact of annotations by LLMs on students' post-editing and translation learning.

1 Introduction

As underlined by the famous quote attributed to Yorick Wilk: "More has been written about MT evaluation than about MT itself" (King et al., 2003). Translation evaluation is an essential but highly challenging task. It relies primarily on two approaches. The first consists in assigning scores that reflect translation quality at different levels — be it a segment, a paragraph, a document, or a system. This type of metric is central to assessing machine translation performance. The second approach, more commonly used in the field of translation studies and translation training, although increasingly used in MT evaluation (see, e.g., Freitag et al. (2021)), involves annotating translations by identifying errors (i.e. words that need to be corrected to improve the translation) and categorising them according to an error typology.

The high cost of human evaluation, whether in terms of time, technical expertise, effort, or financial resources, has driven researchers to explore ways to automate evaluation. While automatic metrics capable of approximating human judgments, with varying degrees of accuracy, have existed for quite some time and continue to improve (Marie et al., 2021), automating error annotation remains a significantly more complex challenge. Until recently, it had attracted little attention and was even considered out of reach.

The launch of ChatGPT in 2022, and more generally the development of LLMs since the 2020s, opened up new possibilities for automating this second type of evaluation. LLMs, initially designed to produce fluid text in natural language, quickly drew the attention of the scientific community for their ability to perform complex tasks, for which they have not been explicitly programmed, taking advantage of their ability to model and manipulate language. Numerous experiments have indeed highlighted the possibility of using LLMs to solve tasks simply by prompting them, i.e. by explaining in natural language how to solve the task at hand. In recent years, pioneering works, reviewed in Section 2, have emerged, highlighting the potential of LLMs, notably ChatGPT, to automate translation evaluation, reflecting a growing interest in integrating

Proceedings of Machine Translation Summit XX Volume 1, pages 190–203, June 23-27, 2025

LLMs into tasks that, until now, have relied mainly on human intervention. Our work follows this dynamic by investigating the possibility of prompting an LLM (here ChatGPT) to annotate MT outputs by identifying and categorising errors. However, we explore this possibility in a new direction: our experiments stand out by focusing exclusively on specialised translation (LSP, language for specific purposes), in opposition to previous research, which primarily considers general language. Specialised translation has considerable economic implications, as it plays a crucial role in industries ranging from law and medicine to technology. Specialised translation also raises additional challenges, both for human translators/evaluators and for MT systems. These include accurate processing of specialised terminology, phraseology, and the management of complex patterns inherent to specialised texts.

We have carried out different annotation experiments using ChatGPT with different prompts and two different MT systems frequently used by professional and non-professional translators alike (DeepL or ChatGPT). Our goals with these experiments are:

- To evaluate the effectiveness of our prompts with ChatGPT when annotating specialised translations (in this case, in the field of natural language processing);
- To analyse ChatGPT's performance in error identification and categorisation, based on an error typology designed for specialised translation evaluation;
- To compare ChatGPT's annotation performances with respect to the MT system being evaluated.

Ultimately, by shedding light on this issue, our work aims to contribute to the creation of hybrid tools, where artificial intelligence and human expertise complement each other to promote more effective learning and teaching, and more accessible self-evaluation or teacher evaluation.

The rest of this work is organised as follows. We will start by reviewing related works in Section 2, before detailing the context and motivations of our work. We will then present our experimental setting in Section 4 and the results of the different experiments we have carried out in Section 5.

2 Related Works

In recent years, the NLP and translation community has shown interest in exploiting LLMs as translators, with promising results. Many even imply that the future of MT is closely linked to LLMs and generative AI (cf. e.g. Lyu et al. (2024); Wang et al. (2023); He (2024); Jiao et al. (2023); Siu (2023)). If LLMs are able to produce translations, they should also be able to assess translations and distinguish between high and low-quality translations. This assumption was the starting point for a group of researchers in NLP and translation studies who, in early 2023, began exploring the ability of LLMs to evaluate translations.

Using LLMs to predict human judgements One of the pioneering works in this field is that of Kocmi et al. (2023), who created the GEMBA metric in zero-shot mode[1], both with and without reference. Their primary goal was to compare the evaluations performed by 9 different GPT models with reference human annotations from WMT'22 (Kocmi et al., 2022) and to observe the level of correlation between the 2 types of evaluation, both at the system and segment level. GPT-based evaluations were carried out with scoring (direct assessment and Scalar Quality Metric SQM) and classification (quality classes) tasks. Their experience on 3 high-resource language pairs shows that, with direct assessment, GEMBA with reference achieves state-of-the-art performance in comparison with other WMT'22 metrics. Without reference, i.e., for quality estimation tasks, GEMBA is the best metric. The best performance is achieved when using GPT-3.5 and higher models, especially GPT-4 (OpenAI et al., 2024).

Using LLMs for Error Annotation Later, Lu et al. (2024) went one step further by using ChatGPT to evaluate translation quality through error analysis (EA) prompting using WMT'20 data (Barrault et al., 2020), again with high-resource languages. Their goal is still to compare the evaluations of ChatGPT (made here in a few-shot and chain-of-thought (CoT) mode) with reference annotations: they asked ChatGPT to identify minor and major errors based on the MQM typology[2] and to score them in order to achieve state-of-the-art

[1]Zero-shot learning is a machine learning paradigm in which a model can make predictions by leveraging semantic knowledge, such as descriptions, rather than relying solely on labeled training examples.

[2]https://themqm.org/

performance at both the system and segment levels. Their results show that their EA metric achieves state-of-the-art performance at the system level, but lags behind other metrics at segment level. However, they show that combining CoT and EA improves evaluation capabilities at the segment level, provided that the prompt includes examples (few-shot).

Another work is that of Fernandes et al. (2023), who created the AutoMQM metric with and without reference, aiming to identify and classify errors according to MQM (interpretable metric) and produce a quality score with the PaLM (Chowdhery et al., 2024) and PaLM-2 models (Anil et al., 2023), but with both high-resource languages using WMT'22 data (Kocmi et al., 2022) and low-resource languages using WMT'19 data (Ma et al., 2019). They aim to show how fine-tuning on human annotation data boosts the performances of LLMs. Their results show that prompted AutoMQM achieves state-of-the-art performance at the system level, but that fine-tuning is necessary to boost performance at the segment level, especially without reference. They also show that adding in-context examples to prompts improves model performance. Experiments with low-resource languages show that LLMs are still underperforming.

With this growing interest in applying error analysis to LLM-based evaluation, Kocmi and Federmann (2023) created the reference-free GEMBA-MQM metric, based on two versions of GPT, aimed at annotating MQM-based errors and evaluating the performance of their metric at system level using data from WMT'22 (Kocmi et al., 2022) and WMT'23 (Kocmi et al., 2023). Their prompting is single-step and three-shot. They show that GEMBA-MQM achieves state-of-the-art performance compared with other metrics without human reference, and also outperforms many metrics with reference.

To the best of our knowledge, the latest work to date is that of Lu et al. (2025) with the MQM-APE metric, aiming to improve the quality of error annotations by 8 open-source LLMs with MQM without reference in order to boost the performance of MQM-APE over other baseline metrics at both system and segment level. They used data from WMT'22 (Kocmi et al., 2022) for high-resource languages and IndicMT (Sai B et al., 2023) for low-resource languages. Their method consists of several steps: (a) MQM-based error annotation by LLMs using the GEMBA-MQM prompt; (b) post-editing by LLMs of annotated segments to determine errors that affect translation quality; (c) checking quality between pairs before and after post-editing to see whether PE improves the original translation. Errors that are not corrected are not counted as errors. The score of the original translation is calculated based on the errors counted after step (b). They then compare MQM-APE and GEMBA-MQM to show that MQM-APE improves performance at both the system and segment level, for high-resource and low-resource languages.

LLMs for evaluating human translations This overview shows the rapid development of this research field within the NLP community and for NLP purposes. However, a few works also focus on the use of LLMs, in particular ChatGPT, for practical purposes, including for translation training. For example, Araújo and Aguiar (2023) used ChatGPT to evaluate translations by taking into account fluency, adequacy and appropriateness, each of these criteria being rated from 1 to 5 by ChatGPT. They compared ChatGPT annotations with reference annotations. The results show a consensus with regard to the lowest-scoring translation, but some variation in the best translations. Still, they show that ChatGPT is a reliable tool for researchers who regularly use MT to translate articles: ChatGPT can be useful for researchers who want to evaluate their machine translated texts, especially as it is an interactive tool offering recommendations, corrections, etc.

Cao and Zhong (2023) used ChatGPT in a pedagogical context. They compared 3 types of feedback for students (teacher feedback, self-feedback and ChatGPT feedback) on the basis of seven linguistic indicators (for lexicon, syntax and cohesion) and by evaluating the final versions after these feedbacks using the BLEU score (Papineni et al., 2002) with reference translations by professionals. They show that for cohesion and syntax, ChatGPT is no more useful than teacher feedback or self-feedback. On the other hand, ChatGPT improves students' lexis more than the other two types of feedback. They therefore suggest adopting a mixed approach for the three types of feedback, combining the capabilities of AI with the more conscious and nuanced feedback of teachers. One of our long-term goals is also to make use of the LLMs' capabilities in a pedagogical context, both by the students themselves and by teachers for evaluation purposes.

3 Context and goals

In this work, we are conducting experiments aimed at assessing the capabilities of LLMs to evaluate the quality of translations using prompting only, but with different motivations and objectives than those of the works outlined in Section 2. We are investigating whether LLMs can identify and categorise errors in a translation, and particularly in specialised translation. In light of the works described in Section 2, it seems appropriate to focus solely on prompting and not consider fine-tuning, since prompt-based evaluation already delivers strong results and, more importantly, the number of languages and domains for which error-annotated corpora of MT exist (especially for LSP translation) is too small for model fine-tuning to be considered a relevant solution.

In this context, given that our goal is to ask an LLM, namely ChatGPT, to identify and categorise errors, it is necessary to rely on an error typology that covers all the issues likely to arise in translations in order to ensure the effectiveness of this approach and the consistency of annotations. The typology we used is based on the MQM typology and on MeLLANGE (Multilingual e-learning in language engineering) (Castagnoli et al., 2011; Kübler, 2008), an annotation framework designed for annotating translations in a translation training context. Even if it is not the main objective of this work, the possibility of using LLMs to identify errors in students' translations offers many interesting prospects, whether for evaluation assistance or to help students in their learning.

The modifications we have made[3] make it possible to adapt these two typologies to the evaluation of specialised translation. This includes, among other factors, a more granular categorisation for terminological errors,[4] in order to account for the complex and domain-specific nature of such texts.

4 Experimental Method

4.1 A prompt for identifying errors in translations

The core of our work is based on the development of a prompt that enables an LLM to identify errors defined in a given typology. Unlike many research efforts in this field which, in line with the way translations are evaluated in the MT community, directly produce a score corresponding to the overall quality of a system or a translation, the prompt we have developed has a dual objective: to precisely locate the words and segments in the translation that are incorrect (the notion of "correctness" being defined by the error typology) and to characterise these errors by assigning them an error type (label) defined by the typology. After several trials and errors, we came up with a prompt whose results on a small set of examples seemed satisfactory enough to be systematically tested on a large scale[5].

Our final prompt is a prompt in French[6], containing the instructions (task requested and its purpose, text type, explanation of attached file, expected output presentation), the error typology with a definition for each type of error, and the text to be annotated along with its source. In addition to the information contained in the instructions, we provide our full annotation manual[7] as an attachment to the LLM. Given the large amount of text in the prompt, we used the prompt chaining technique[8] (Ekin, 2023) and zero-shot mode, as no examples are included in the instructions. Although each text in the corpus was translated at the document level, and not by sentence, we opted for sentence-level alignment when using ChatGPT to annotate errors, in order to minimise the volume of densely-packed information to be processed by the model. The full prompt is given in Figure 5 (Appendix B).

Note that, with the exception of one sentence specifying the type of text translated (abstracts of research articles in NLP), our prompt does not contain any instructions specifically relating to the text type or to the (highly) specialised domain. Therefore, although we have not specifically tested this aspect, it seems likely that the results we report in this work can also be applied to other types of text.

[3]The full typology is described in Figure 4 in Appendix A.

[4]As shown in Figure 4, the error category relating to terminology contains 10 error subtypes.

[5]The prompt used was designed by a translator with prior experience in translation evaluation, but no extensive training in prompting and NLP, highlighting the fact that for this task and for the purposes at hand, it seems more appropriate to rely on an expert in translation evaluation rather than an expert in prompting, especially given the effectiveness of the prompt.

[6]We carried out preliminary experiments with an English prompt, and the results indicated no significant difference between the English and the French prompts, although the latter performed slightly better on the sample tested.

[7]The annotation manual is a 50-page document designed to guide an evaluator in annotating translations according to our error typology. It provides general annotation guidelines, a full explanation of the typology, a definition and various examples for each error type.

[8]The prompt chaining technique involves linking multiple prompts together sequentially, where the output of one prompt becomes the input for the next one, enabling complex, multi-step reasoning or task completion.

In our experiments, we experiment with two variations of this prompt and use it to identify errors in translations produced by different mainstream translation systems.

4.2 Reference human annotations

To evaluate the ability of an LLM to identify errors in the translation of a specialised text, we built a corpus comprising source documents (abstracts of NLP research articles in English from the HAL open archive[9]), their translation in French by different MT systems used by both the general public and professional translators (namely DeepL and ChatGPT[10]) and an annotation of these translations by a human expert (a professional translator with extensive experience in evaluating translations and using our error typology) who identified the errors contained in these MT outputs. In this context, "annotation" refers to the manual identification and labelling of errors in a translation, where the annotator identifies incorrect segments and assigns one or more error categories based on our predefined typology. The annotated translations contain error spans and error type labels (occasionally several possible labels) for each error.

In the end, our corpus[11] is divided into two sub-corpora: a) a sub-corpus of 35 source texts translated by DeepL with the annotated translations based on the error typology (10,500 words[12]), and b) a sub-corpus of 25 source texts translated by ChatGPT with the annotations based on the typology (7,431 words).

Figure 1 shows an example of the annotation produced by our expert. In the first sub-corpus, the expert identified 399 errors (an average of 11.4 per document); the errors ranged from 2 to 81 characters (average: 15 characters), and had between 1 and 6 possible labels (average: 2.3). For the second sub-corpus, the expert identified 193 errors (on average 7.7 per document); the errors ranged from 2 characters to 103 characters (average: 22 characters), and had between 1 and 4 possible labels (average: 2.1).

[9] https://hal.science/

[10] Here is the prompt (translated in English) we used to translate the texts with ChatGPT: "You are a translator who specialises in translating research articles on natural language processing. Translate the following text into French, respecting the structure of the original text and not omitting any elements."

[11] Our corpus of French translation reference annotations with English source texts is available here: https://doi.org/10.34847/NKL.52E571A3

[12] To count the number of words, we naively tokenised our corpus using spaces.

4.3 Evaluating the evaluations

In order to automatically evaluate the performance of our prompts in detecting errors identified by the expert translator, we use the standard recall and precision metrics commonly employed in NLP to assess error detection systems. Precision measures the proportion of errors identified by an LLM with our prompt that are actually correct. It is calculated as the ratio of true positives (correct corrections) to the total number of corrections made (true positives + false positives). Precision reflects the system's ability to avoid making incorrect corrections. Conversely, recall gauges our prompts's capacity to pinpoint all errors present in a text. It is calculated as the ratio of true positives to the total number of actual errors in the corpus. This number is given by the sum of the number of true positives (the number of errors correctly identified) and of false negatives (the number of errors "missed" by the model).

A high precision indicates that our system makes very few incorrect corrections, but it does not necessarily mean that all existing errors are detected. Conversely, a high recall shows that the system identifies most of the errors in a text but might introduce many false corrections, leading to lower precision. To easily compare the performance of the different prompts we consider, we use the standard F_1 score, which combines recall and precision into a single number to compare the performance of two systems.

Defining these three metrics involves determining whether an error identified by the expert corresponds to a predicted error. However, this is not always straightforward, as the definition of an error can be subjective and open to interpretation: the decision of whether to include a word in the definition of an error can vary between annotators. For practical reasons, we decided to consider an error in the reference annotation as correctly identified by the LLM if the error shares at least one character with a predicted error.[13] This decision is based on the assumption that even a single shared character is enough to draw a translator's attention to the area with a potential issue.

With these definitions, precision P and recall R are simply defined as :

$$P = \frac{\text{number of errors correctly identified}}{\text{number of predicted errors}} \quad (1)$$

[13] We have also ensured in our evaluation that a reference error is not associated with two different predicted errors.

Les contes de fées, les contes du peuple LA-TL-INS, LA-TL-ING et plus généralement les histoires d'enfants TR-DI, LA-SY-PR, LA-SY-GNC, LA-TL-INS, LA-TL-ING ont récemment attiré la communauté du Traitement Automatique des Langues (TAL). A ce titre LA-HY-PU très peu de corpus existent, et les ressources linguistiques manquent. Le travail présenté dans cet article vise à combler la lacune LA-UR, LA-TC-CE, LA-TC-CN, LA-SY-DET, LA-ST-AW en présentant un corpus annoté syntaxiquement et sémantiquement. Elle LA-IA-GE, LA-UR, LA-TC-CE, LA-TC-CN se focusse TR-SI-UT, TR-SI-TL, LA-TL-ING sur l'analyse linguistique d'un corpus de contes de fées et fournit une description des ressources syntaxiques et sémantiques développées pour l'extraction des informations LA-TL-INS, LA-SY-DET, LA-SY-PR.

Figure 1: Example of a human reference annotation: each error is identified by its span (text written on an orange background) and one or more labels (in subscript).

and

$$R = \frac{\text{number of errors correctly identified}}{\text{number of errors in reference}} \quad (2)$$

In our evaluation, precision and recall will be calculated at the level of each document, enabling a fine-grained analysis of the system's performance across individual texts. The results will then be averaged over all documents, meaning the reported numbers correspond to macro-recall and macro-precision.

In several cases, an error can be tagged with more than one error label; this is the case, for example, with terminological errors, which can distort the meaning of the message (terminological error + content transfer error). In order to assess a prompt's ability to correctly categorise errors, we also report, for each experiment, the proportion of correctly identified errors whose predicted label matched at least one reference label, since the model predicts only one label for each error. Asking the LLM to predict multiple labels would make the task too complex, both for the LLM and our meta-evaluation.

4.4 Experiments

So far, three different experiments have been carried out. The first experiment, denoted "`long prompt`" in the following, consisted in having 35 MT outputs in French from DeepL evaluated by ChatPT (OpenAI et al., 2024)[14] with the prompt described in Section 4.1. The second, denoted "`short prompt`", with the same 35 MT outputs, involved testing a shorter and less information-laden version of the prompt, i.e. removing the definitions of each type of error; this follows suggestions made by Lu et al. (2024), who recommend against providing error descriptions in detail. Finally, the last experiment involved ChatGPT evaluating 25 MT outputs of other source texts it had generated itself.

[14]We used the version of ChatGPT that relies on GPT-4o.

	MT System	
	DeepL	ChatGPT
# texts	35	25
# gold errors	399	193
long prompt		
# pred. errors	384	224
precision	0.792 $\pm$ 0.0396	0.47 $\pm$ 0.0989
recall	0.653 $\pm$ 0.0488	0.57 $\pm$ 0.107
F_1	0.707 $\pm$ 0.0393	0.496 $\pm$ 0.0933
% correctly labeled	64.1 %	45.3 %
short prompt		
# pred. errors	417	—
precision	0.745 $\pm$ 0.0575	—
recall	0.671 $\pm$ 0.0505	—
F_1	0.702 $\pm$ 0.0531	—
% correctly labeled	46.9 %	—

Table 1: Results achieved by our different prompts on the two corpora we consider. "# gold errors" represents the number of errors found by the expert annotator, "#pred error" represents the number of errors predicted by our system. We have computed the 95% confidence intervals for the different scores we consider using the bca bootstrap method of Efron and Tibshirani (1993).

The primary aim of these experiments is, firstly, to see whether ChatGPT can perform annotation tasks on specialised translations with our error typology. Next, we aim to understand the strengths and weaknesses of this model, especially in terms of error identification and categorisation. We also intend to see whether defining each error in the prompt influences the quality of annotations and whether its capabilities vary according to the source of the MT output (DeepL or its own translations).

5 Results

As explained in Section 4.3, we measure the ability of the different prompts considered to correctly

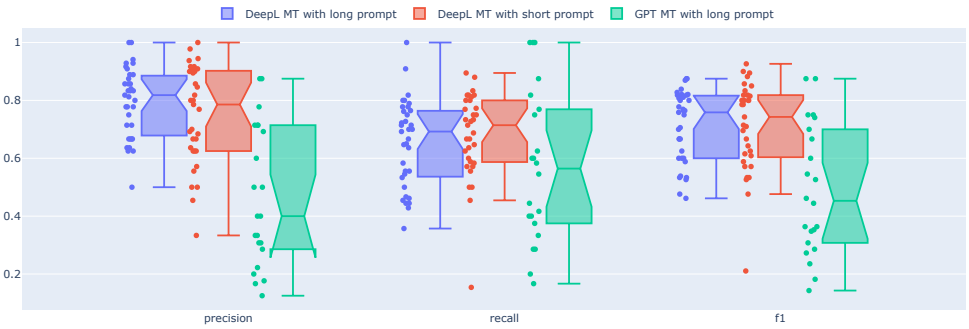


Figure 2: Distribution of precisions, recalls and F_1 scores across documents for the different prompts we consider.

identify translation errors by evaluating recall, precision and F_1 score on two corpora of translations generated by different MT systems. The results obtained are summarised in Table 1. For the sake of clarity, we have also reported in this Table the total number of gold errors (i.e. errors identified by an expert) in our corpus, the number of predicted errors and the percentage of labels that are correctly identified. Figure 2 also shows the distribution of the various scores obtained to enable a more detailed analysis of the performance of the prompts. In the remainder of this Section, we will detail the results achieved for each experiment.

Annotations of DeepL MTs Using ChatGPT with the *long* prompt to identify and categorise errors in DeepL MT outputs shows promising results.

For all errors identified in reference human annotations, ChatGPT identifies between 6 and 7 out of 10. The model also seems to perform very well when it comes to categorising errors based on the error typology, managing to accurately categorise around 65% of them. If a sentence contains no errors — which does happen —, ChatGPT occasionally acknowledges the fact that there are no errors in the sentence, but it can also over-annotate the translation by detecting errors that are not actually errors (what we call "false errors" here). On average, in an annotated text, ChatGPT identifies between 1 and 2 "false" errors, i.e. errors that are not identified as errors in the reference annotation. The number of false errors varies between 0 and 5 per text. These false errors represent 14.47% of the errors annotated by the model.

Although the average F_1 score (0.71) indicates a satisfactory overall performance, the dispersion of scores (Figure 2) shows that the model can react unpredictably to different texts: depending on the document, precision may vary from 1.0 to 0.5 and recall from 1.0 to 0.35. This variability could reflect sensitivity to differences in the complexity or nature of the errors to be identified, making performance occasionally more random depending on the case. These rather unpredictable performances of ChatGPT have already been pointed out by the scientific community (see, for example, Siu (2023)). However, it does call into question the practical interest of the model: it is unlikely that a translator would use such a system to identify errors if they were randomly wrong.

Using a shorter prompt by removing the definition given to each type of error in the instructions (see § 4.1) shows similar results. The system's ability to identify errors is more or less the same: the overall F_1 score is also around 0.70. Whereas, with the full prompt, ChatGPT performed better in error categorisation than in error identification, the opposite happens with the short prompt. In fact, it identifies almost 7 out of 10 errors. On the other hand, around 5 out of 10 errors are incorrectly categorised.

This drop in error categorisation performance comes as no surprise, since the prompt no longer contains the definitions of each type of error. However, it is more surprising to see that the removal of this information has only a (very) slight impact on the system's ability to identify errors, suggesting that ChatGPT's ability to identify translation errors is not linked to the information it has extracted from the prompt, but only to the knowledge it has acquired during its training or to the knowledge it acquires from the attached annotation manual.

As far as false errors are concerned, the average here is 1.71, and per text, the number of false errors varies between 0 and 7. False errors account for

17.86% of all errors annotated by ChatGPT with the short prompt. For this test, recall is slightly higher than in the first experiment (0.67 compared with 0.65), and we observe a slight loss of precision, reaching 0.75.

It is also interesting to note that, as shown in Figure 2, removing the definition for each error type from the prompt significantly increases performance variability: In contrast to the performance of the long prompt, where the lowest precision was 0.500, here several texts (6) show clearly low scores, highlighting specific difficulties or cases where the model performs less well. This comparison highlights a more marked uncertainty in the reliability of the model's evaluations on this set of texts with the short prompt.

Annotations of ChatGPT MT outputs ChatGPT annotations of its own 25 MT outputs with the full prompt show particularly weak results compared to the two previous experiments.

Table 1 shows that when annotating its own MT outputs, ChatGPT identifies only about half of the errors contained in the reference annotations. Well-categorised errors are also below 50%. The rate of false errors per text doubles or even triples compared with the two previous experiments, reaching almost 5 false errors per annotated text. They account for 55.02% of all errors identified by the model, i.e. more than half, and range from 0 to 14 per text.

The average overall F_1 score is significantly lower than it was in the two previous experiments, dropping to 0.496. In terms of recall and precision, the results are no better, with a recall of 0.57 and a precision of 0.47.

Figure 2 shows a low average score and high variability, reflecting limited performance in this particular setting. This can be explained by the fact that ChatGPT evaluated its own machine translated texts, a task that seems to raise specific challenges. The large number of scores below 0.5 suggests that the model struggles to identify and categorise its own errors in a systematic way, probably due to implicit bias or a lesser ability to step back from its own productions. This contrast with the other evaluation scenarios highlights a weakness in the model's self-evaluation.

6 Discussion

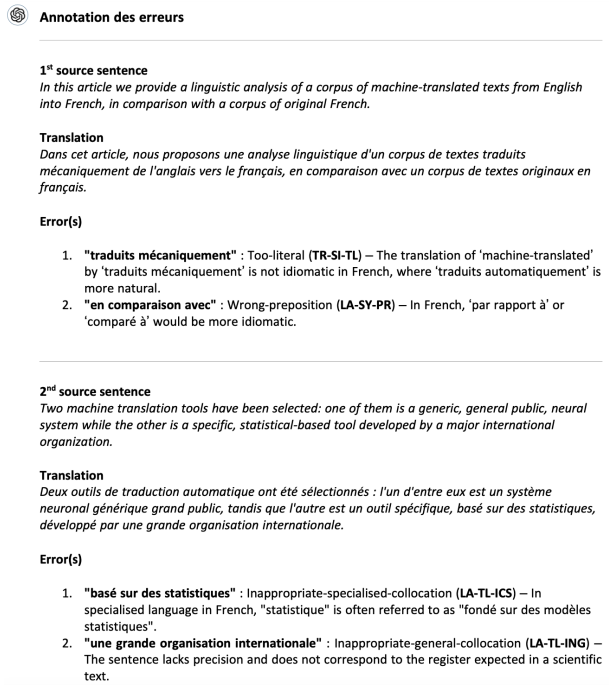
Results achieved with the full prompt for DeepL MT annotations show satisfactory performances of ChatGPT, with an average F_1 score of 0.707 and a fairly strong capability to identify and categorise errors (about 65%). In contrast to what has been claimed by the community (see e.g. Lu et al. (2023)), our experiments show that providing a detailed definition for each type of error in the prompt slightly improves the model's performance, in particular for error categorisation, which was below 50% with the short prompt and reached 65% with the definitions in the prompt.

Despite these encouraging results, however, the variability in scores, particularly with the short prompt and with ChatGPT's MT outputs, suggests that ChatGPT's effectiveness with these parameters is not consistent. Indeed, a large number of texts in these two experiments have very low recall, precision and F_1 scores.

Finally, ChatGPT shows considerable limitations in self-assessment, with significantly lower performance when annotating its own translations (F_1 score of 0.496). This result highlights a possible bias and the model's lack of ability to objectively evaluate its own productions, which justifies our motivations to set up annotation experiments with other open-source LLMs, as have Fernandes et al. (2023) and Lu et al. (2025), demonstrating that comparable — or even better — performance can also be achieved with open-source LLMs.

In order to further assess the relevance and usability of ChatGPT's outputs, we provide an example of an individual annotation[15] performed by the model (see Figure 3). This annotation by ChatGPT clearly shows that the LLM annotates the errors as instructed in the prompt, respecting the requested output format (sentence by sentence), giving the error span, the error category, the associated label and an explanation. However, we did not directly request explanations for each annotated error in our prompt, but the model seems to do so systematically. For these sentences, the explanations the model provides make sense. These explanations highlight the potential usability of these experiments with ChatGPT: since the LLM provides explanations and possible solutions for each potential error, it seems reasonable to consider conducting other experiments in a real-life classroom setting using ChatGPT's annotations (see Section 7).

[15]Figure 3 represents an annotation produced by ChatGPT. However, this is not the output we considered for calculating comparison scores with the reference annotations. To calculate these scores, we asked ChatGPT to convert its annotations into a table that could be used and analysed automatically.



Annotation des erreurs

1st source sentence
In this article we provide a linguistic analysis of a corpus of machine-translated texts from English into French, in comparison with a corpus of original French.

Translation
Dans cet article, nous proposons une analyse linguistique d'un corpus de textes traduits mécaniquement de l'anglais vers le français, en comparaison avec un corpus de textes originaux en français.

Error(s)

1. **"traduits mécaniquement"** : Too-literal (**TR-SI-TL**) – The translation of 'machine-translated' by 'traduits mécaniquement' is not idiomatic in French, where 'traduits automatiquement' is more natural.
2. **"en comparaison avec"** : Wrong-preposition (**LA-SY-PR**) – In French, 'par rapport à' or 'comparé à' would be more idiomatic.

2nd source sentence
Two machine translation tools have been selected: one of them is a generic, general public, neural system while the other is a specific, statistical-based tool developed by a major international organization.

Translation
Deux outils de traduction automatique ont été sélectionnés : l'un d'entre eux est un système neuronal générique grand public, tandis que l'autre est un outil spécifique, basé sur des statistiques, développé par une grande organisation internationale.

Error(s)

1. **"basé sur des statistiques"** : Inappropriate-specialised-collocation (**LA-TL-ICS**) – In specialised language in French, "statistique" is often referred to as "fondé sur des modèles statistiques".
2. **"une grande organisation internationale"** : Inappropriate-general-collocation (**LA-TL-ING**) – The sentence lacks precision and does not correspond to the register expected in a scientific text.

Figure 3: Example of annotation by ChatGPT. This figure shows the sentence-by-sentence annotation performed by ChatGPT, which identifies the error, categorises it, assigns a label and provides explanations and solutions for improvement. The initial output of ChatGPT is in French, since the prompt provided is written in French. For the purposes of this article, we have translated it into English.

7 Conclusion

This study explored the use of ChatGPT for annotating MT outputs based on a customised error typology adapted to our specific needs in a specialised translation training setting. The annotations generated by the model were compared with reference human annotations to evaluate its ability to identify and categorise errors in a translation generated by DeepL or ChatGPT. Initial results are encouraging, particularly with external machine translations, where ChatGPT identified and categorised most errors with reasonable accuracy, in particular with the long prompt containing the definition for each type of error. However, its performance was far less reliable when evaluating its own translations.

Another key finding from our experiments is that, given the lack of a significant difference in error identification between the full and short prompts, it seems reasonable to suggest that the structure and degree of detail of the prompt does not have a major impact on ChatGPT's performance in this annotation task. This could indicate that ChatGPT is performing its annotations efficiently even without detailed instructions (without a definition for each error), relying more on its knowledge acquired during training rather than on the specifications of the prompt. However, this also suggests that while LLMs can rely on their pre-trained knowledge to identify errors, their ability to categorise these errors correctly benefits from clear, structured instructions and definitions.

Future Work Future experiments will extend this research to open-source LLMs, focusing on their potential to provide annotations of comparable or superior quality. These models, with greater transparency, will be evaluated not only for their accuracy and capabilities in annotating translations but also for their ease of integration into automated workflows for translation quality assessment.

Ultimately, our aim is to test the effectiveness of this automated evaluation by LLMs in a practical context of translation training. Firstly, we intend to optimise the human evaluation process. Specifically, with teachers annotating students' translations, we aim to examine whether the use of annotations generated by LLMs can reduce the cognitive effort associated with the annotation process. Additionally, we intend to carry out experiments with translation students and test whether the use of LLM annotations help them improve the quality of their MT post-editing. Our aim is to test our prompt with other domains, notably earth and planetary science.

Limitations

ChatGPT, as an OpenAI proprietary model, has some limitations that need to be taken into account in our experiments. The lack of transparency regarding its training data, the uncertainties associated with its availability in the future and the fluctuations in its performance over time make it difficult to assess its capabilities in a rigorous and reproducible way. These issues have been highlighted by other researchers, notably Chen et al. (2024), who observed significant variations over the course of 2023. That being said, ChatGPT remains a mainstream tool that is used by many translators in their day-to-day work. Therefore, we believe that it would be relevant to evaluate it. However, we intend to conduct similar experiments with other open-source LLMs, which have already demonstrated state-of-the-art performance. These models offer greater transparency and full control over the versions used, which is essential to guarantee traceable and reproducible results.

Acknowledgments

This research was funded by the French Agence Nationale de la Recherche (ANR) under the project MaTOS - "ANR-22-CE23-0033-03".

References

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. *Preprint*, arXiv:2305.10403.

Sílvia Araújo and Micaela Aguiar. 2023. Comparing chatgpt's and human evaluation of scientific texts' translations from english to portuguese using popular automated translators notebook for the simpletext lab at clef 2023.

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.

Siyi Cao and Linping Zhong. 2023. Exploring the effectiveness of chatgpt-based feedback compared with teacher feedback and self-feedback: Evidence from chinese to english translation. *Preprint*, arXiv:2309.01645.

Sara Castagnoli, Dragos Ciobanu, Kerstin Kunz, Natalie Kübler, and Alexandra Volanschi. 2011. Designing a Learner Translator Corpus for Training Purposes. In Natalie Kübler, editor, *Corpora, Language, Teaching, and Resources : From Theory to Practice. Bern: Peter Lang*, volume Etudes Contrastives of *Corpora, Language, Teaching, and Resources : From Theory to Practice. Bern: Peter Lang*, pages 221–248. Peter Lang.

Lingjiao Chen, Matei Zaharia, and James Zou. 2024. How Is ChatGPT's Behavior Changing Over Time? *Harvard Data Science Review*, 6(2). Https://hdsr.mitpress.mit.edu/pub/y95zitmz.

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, Sashank 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. 2024. Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1).

Bradley Efron and Robert Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, London.

Sabit Ekin. 2023. Prompt engineering for chatgpt: A quick guide to techniques, tips, and best practices.

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. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

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).

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. *Preprint*, arXiv:2301.08745.

Margaret King, Andrei Popescu-Belis, and Eduard Hovy. 2003. FEMTI: creating and using a framework for MT evaluation. In *Proceedings of Machine Translation Summit IX: Papers*, New Orleans, USA.

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 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.

Natalie Kübler. 2008. MeLLANGE Final Report. Intern report, Université Paris Diderot.

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. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint*.

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.

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. A paradigm shift: The future of machine translation lies with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.

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, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation re-

search: A meta-evaluation of 769 papers. 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 7297–7306, 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, 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, Ben 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, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha 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, Johannes 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, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz 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, Mateusz 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 Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario 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 Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll 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, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics.

Sai Cheong Siu. 2023. Chatgpt and gpt-4 for professional translators: Exploring the potential of large language models in translation. *SSRN Electronic Journal*.

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 Error typology

B Full prompts

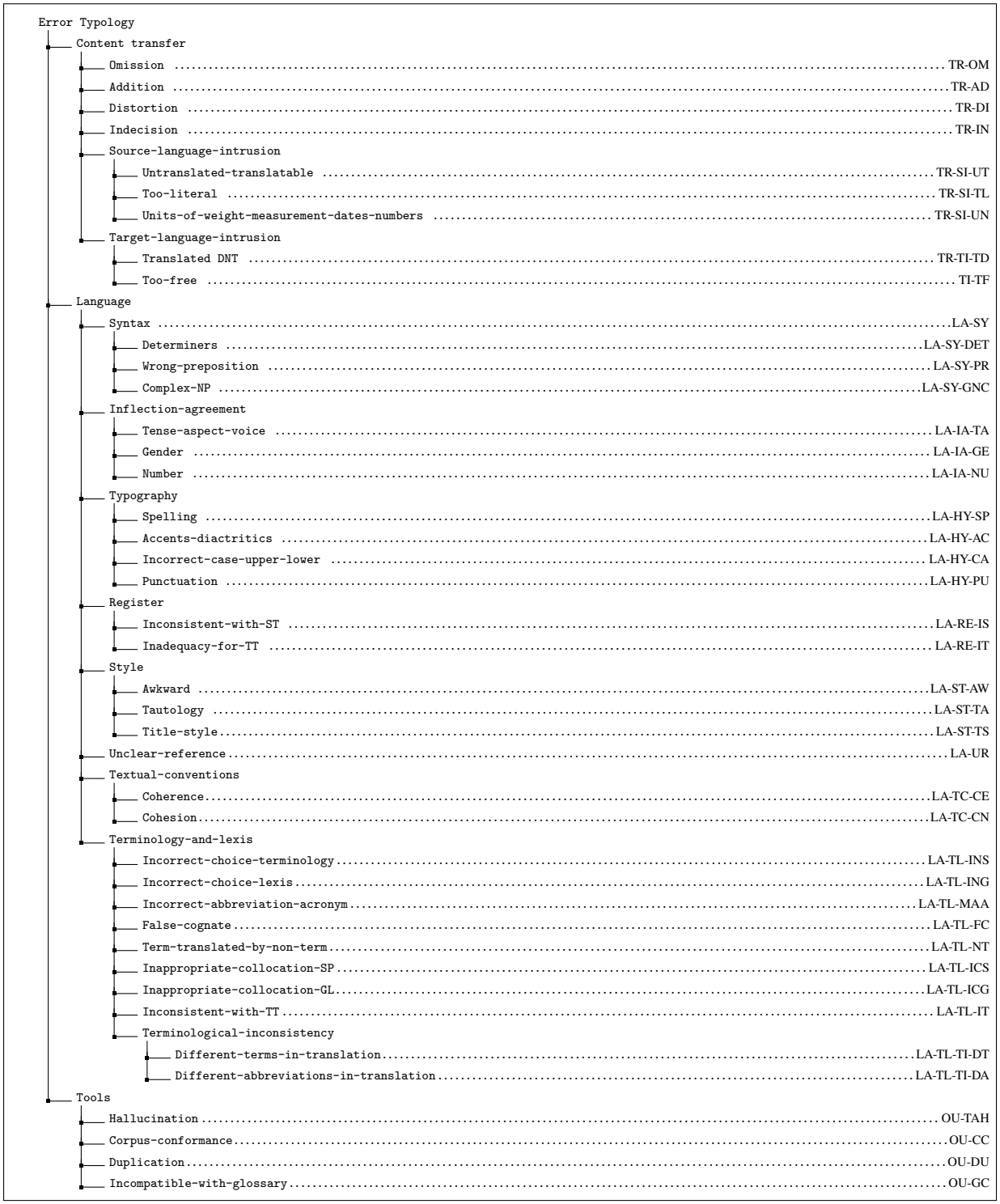


Figure 4: The error typology used in our experiments.

1. Tâche : annoter une traduction
   Objectif : repérer des erreurs sur la base d'une typologie d'erreurs que je te fournis.
   Type de texte : résumé d'article scientifique dans le domaine du TAL
   Fichier joint : MANUEL D'ANNOTATION, qui contient des explications plus détaillées et des exemples des types d'erreurs que je vais te fournir ci-dessous.
   Présentation de la sortie :
   - 1re phrase source
   - 1re phrase cible dans la traduction
   - liste les erreurs
   Etc. jusqu'à la fin de la traduction

   Je vais te donner la typologie d'erreurs.

2. Typologie d'erreurs à suivre méticuleusement : veille à utiliser les types d'erreurs présents et n'en invente aucun. De même, respecte les codes liés à chaque type d'erreur à la lettre ; ne prends donc aucune liberté.
   Explication de la typologie : elle est divisée en 3 grandes catégories d'erreurs : les erreurs de transfert de contenu (erreurs altérant le sens du message ou entravant sa compréhension), les erreurs de langue, et les erreurs liées aux outils ou à leur maîtrise.
   Voici la typologie :
   1. Transfert-contenu (GRANDE CATÉGORIE, NE PAS UTILISER)
   1.1. Omission_TR-OM
   * Une omission se produit lorsqu'il manque, dans la traduction, une idée qui est présente dans le texte source. Il ne faut pas confondre omission et implicitation. Une omission a lieu sans réelle raison valable, alors qu'une implicitation est un moyen d'éviter une surtraduction.
   1.2. Rajout_TR-AD
   * À l'instar de la différence entre omission et implicitation, on peut souligner une différence de nuance entre le rajout et l'explicitation. L'ajout est considéré comme une erreur, alors que l'explicitation peut s'expliquer par le fait que le traducteur ou le post-éditeur souhaite éviter la sous-traduction.
   ... jusqu'au bout de la typologie ...

   - Prête attention à tous les aspects, autant le transfert de contenu que la langue et la terminologie et les erreurs liées aux outils.
   - Si tu as besoin d'exemples, réfère toi au manuel d'annotation en pièce jointe.

   Je vais te donner la traduction à évaluer avec son texte source.

3. Voici le texte source et sa traduction à annoter :
   (source text)
   (target text)

   PROCÈDE À L'ANNOTATION. Attention, n'annote QUE les erreurs, pas des améliorations ou suggestions ! Il peut y avoir plusieurs erreurs dans une même phrase.

Figure 5: Prompt used on GPT-4o

Name Consistency in LLM-based Machine Translation of Historical Texts

Dominic P. Fischer and **Martin Volk**
University of Zurich
Department of Computational Linguistics
dominicphilipp.fischer@uzh.ch

Abstract

Large Language Models (LLMs) excel at translating 16th-century letters from Latin and Early New High German to modern English and German. While they perform well at translating well-known historical city names (e.g., *Lutetia* → Paris), their ability to handle person names (e.g., Theodor Bibliander) or lesser-known toponyms (e.g., *Augusta Vindelicorum* → Augsburg) remains unclear. This study investigates LLM-based translations of person and place names across various frequency bands in a corpus of 16th-century letters. Our results show that LLMs struggle with person names, achieving accuracies around 60%, but perform better with place names, reaching accuracies around 90%. We further demonstrate that including a translation suggestion for the proper noun in the prompt substantially boosts accuracy, yielding highly reliable results.

1 Introduction

Large language models (LLMs) have shown remarkable capabilities in translating 16th-century texts in Latin and Early New High German into modern-day German or English, surpassing other Machine Translation systems by a considerable margin (Volk et al., 2024a,b). LLMs also excel at translating well-known proper nouns, be they current or historical names. However, the extent to which LLMs accurately translate lesser-known person names and place names (toponyms), is underexplored. Proper nouns are important content words; mistranslations can have far-reaching consequences. If, as it occurred in our experiments, *Vallistellinensi* is mapped to the Valais region in Switzerland instead of Italy's Valtellina, or *Bayonensis* to Basel instead of Bayonne (France), the differences amount to distances of several hundred kilometers!

In this study, we investigate LLM-based translation of person names and place names (countries, cities and regions) in Latin and Early New High (ENH) German from our 16th-century letter corpus into modern German. The translation involves normalizing inflected forms into base forms (e.g. *Bulingero* → Bullinger) but also mapping historical names to their modern-day equivalents (e.g. *Tiguro* → Zürich).

This study is part of a project to translate and provide access to the letter correspondence of the Zurich reformer Heinrich Bullinger (1504-1575)[1]. The corpus consists of 12,000 letters, three quarters of which are in Latin, most of the others in ENH-German (Volk et al., 2022; Ströbel et al., 2024). The corpus is annotated with manually checked person and location names.

The following sentence, with three person names and four place names, exemplifies the subject matter. Note that the GPT-translation splits the long input sentence into two, which helps readability.

1. De synodo tamen **Saxonica** nihil accepi hactenus et ne **Argentorati** quidem (quo his diebus in **Brisgaudiam Alsatiamque**, illuc a **marchionis Ernesti Badensis** filia et vidua, huc vero a **barono a Rapoltzstein** vocatus ex occasione me contuli) quicquam huius intellexi neque eo **d. Marpachius** abiit, quem docentem audivi. [letter 2292 in the corpus]

   GPT-4o: Regarding the **Saxon** synod, however, I have received no news so far, and even in **Strasbourg** (to which I recently traveled on occasion, called to **Breisgau** and **Alsace**, there by the daughter and widow of **Margrave Ernest of Baden**, and here by the **Baron of Rapoltsweiler**), I learned nothing about it. Nor has **Mr. Marbach** gone there, whom I heard teaching."

To evaluate LLM translation performance across

[1] https://www.bullinger-digital.ch/

Proceedings of Machine Translation Summit XX Volume 1, pages 204–219, June 23-27, 2025

a range of name frequencies, we selected a balanced sample of names: the 10 most frequent, 10 of the least frequent (with a minimum of 10 occurrences), and 10 randomly chosen names from intermediate frequency bands. These names appear in many forms, up to 200, as in the case of Heinrich Bullinger himself: *Huldrice, Heinricho, Heimrych, Heilrich, ..., Bullingerus, Bulliger, Bulingero,* etc. For each sampled name, we extracted one sentence for each different form of that name, as well as the context of the three preceding and three following sentences.

We tested with two settings: without context (just the sentence that contains the name) and with a context window of the three sentences before and after. Within these two settings, we explored two different prompting strategies to guide the translation of proper nouns. The two strategies involved including or leaving out historical background information in the prompt (i.e., where the sentence to be translated is taken from), and adapting the system prompt or keeping it generic.

Additionally, we investigated how marking the target proper noun affects the translation quality. This step is interesting in that it might attract the LLM's focus to a particular word or word sequence, and that it structures the output automatically. This facilitates evaluation and postprocessing, since the proper nouns can be easily extracted.

Finally, we evaluated the translation quality if a translation suggestion for the name is included in the prompt.

By evaluating and comparing these approaches, we assess the abilities of LLMs in translating proper nouns and how contextual and formatting cues influence their performance. We also examine the reasons for the discrepancies in translation quality between different names. This work provides insights into how LLMs address complex translation challenges that involve historical texts and proper nouns.

2 The 16th-Century Letter Corpus

Our corpus of 16th-century letters consists of the correspondence of the Zurich reformer Heinrich Bullinger. It encompasses around 12,000 documents, which include 3,100 letters professionally edited by the Institute for Swiss Reformation Studies[2] and an additional 5,400 manually transcribed letters. We have automatic transcripts for most of the remaining 3000+ letters, whereby our handwritten text recognition has a character error rate of around 8%. We ignore these letters in the current study.

The edited part of the corpus has been published in 20 printed volumes (Gäbler et al., 1973–2022), each of which has an index with manually curated person and place names. The index entries for a specific person or place point to the pages in the book where the name appears, but the name itself is not explicitly marked within the printed text.

We used these indices to automatically mark the names in the digital versions of the letter summaries, the letter texts, and the footnotes, initially on the specified pages only. This mapping was limited by the fact that the indices contained the names in standardized form (e.g. the city name *Antwerpen*), which made it difficult and partly impossible to detect all inflected forms and spelling variations of a given name (e.g. *Antwerpia, Antwerpię, Antorff*). All names that we marked in this step come with a unique project-internal identifier, which is linked to Wikipedia or to the GND-database.

In a second step, we copied the assigned identifier of marked names to other occurrences in the same letter where they could be unambiguously assigned. We also applied the annotations to the 5,400 transcriptions. In a third step, we trained a name recognizer on these data to spread the annotation (person and place name tags without linking) to all unmarked name mentions in the 8,500 letters.

In order to improve the name annotation quality, a citizen science campaign contributed by checking person and place names, with volunteers annotating the names and linking them to the corresponding entities. This enriched the corpus with several 10,000 person and place names, enabling us to conduct a comprehensive analysis of proper noun translation. Currently, the 8,500 letters, the roughly 3100 manually written summaries and 75,000 footnotes are marked with 202,000 person name tags and 156,000 place name tags, out of which 188,000 person names (5924 unique ids) and 150,000 place names (2950 unique ids) are linked.

The documents retain historical characters such as ę, ů, å, ǒ, reflecting the orthographic conventions of the period. Abbreviations commonly found in the letters have been expanded by editors and transcribers e.g. '*Frid[olin] Schůler*')[3].

[2] https://www.irg.uzh.ch/

[3] Our experiments showed that such square brackets did not have a discernible effect on translation quality, which is why we left them as is.

The letters constitute a rich historical resource, shedding light on politics, theological debates, regional and European news, education, and family matters. They are part of Bullinger's vast correspondence, whose network extended from Zurich across Europe, reaching as far as Denmark, England, and Lithuania.

The current study focuses on the Bullinger letter exchange, but its results are relevant for the many other letter collections of the same period which amount to more than 100,000 letters in Latin and ENH-German. This applies, for instance, to the collection of the Theologians' Correspondence in the Southwest of the Empire in the Early Modern Period (1550-1620)[4]. For an overview see Hotson and Wallnig (2019).

3 Related Work on Named Entities in Machine Translation

The topic of named entities in machine translation has been addressed repeatedly. For an early paper, see Hirschman et al. (2000), who propose "name translation" as a specific MT evaluation task. More recently Mota et al. (2022) report on "fast improvements" for handling named entities in machine translation by implementing named entity recognition as a separate pre-processing step. Similarly, Zeng et al. (2023) propose an "extract-and-attend" approach to improve neural MT performance between English, Russian and Chinese, that require at least transliteration of the names because of the different scripts.

For translations between languages with the Latin alphabet, Macketanz et al. (2022) report that the "categories with the highest performance (above 90%) were [...] named entities & terminology" when testing various machine translation systems against an English - German test suite that covers many linguistic phenomena.

However, if the target language requires inflection of the names, then translation challenges still arise. One such case is Icelandic. Ármannsson et al. (2024) argue that machine translation of person and place names from English to Icelandic is far from perfect. Interestingly, place names proved to be more difficult than person names in their experiments. Le et al. (2023) show that named entity recognition improves MT for Inuktitut to English.

The central issue is "Lexical Cohesion: The same named entity must be translated consistently across the current sentence and context sentences" (Jin et al., 2023). The ultimate goal is transcreation with cultural adaptation. By integrating information from a multilingual knowledge graph into neural MT Conia et al. (2024) obtained huge improvements for name translation across 10 language pairs.

Although there is previous work on named entity recognition for Latin (Erdmann et al., 2016), we found no paper on named entities in MT for Latin, nor for Early New High German. We are breaking new ground in systematically evaluating names in machine translation from these historical languages to modern languages with LLMs.

4 Methodology: Translating from Latin & Early New High German into German

4.1 Name Selection

To evaluate the translation performance of LLMs on proper nouns, we selected a balanced sample of names. First, we computed the frequencies over the assigned name ids and filtered out names that occur less than 10 times in our corpus. Then, we sampled the 10 most frequent person and place names in the corpus, 10 of the least frequent names, and 10 randomly selected names from intermediate frequency bands. We list the selected persons and places, their frequency in the corpus and example sentences in the appendix. In our test set of person names, slightly less than half of the names are first name + last name combinations, and slightly more than half are single names, i.e. either first name or last name. About 8% of the names contain abbreviations, and another 11% abbreviated names were expanded by editors. Apart from the emperor Karl V., it does not include names of dignitaries, nor does it include special names such as discontiguous names, which we discuss in sections 6.1 and 6.2.

Since we do not distinguish place names from place adjectives in our annotation, both may occur in our test set. For example, the test set covers both *Gallia* (France) and *Gallus* (French). Town names may appear in full form such as *Augusta Vindelicorum*, or shortened to a frequently used part such as *Augusta* (both referring to Augsburg).

For each sampled name, we extracted one sentence for each distinct form of the name, along with the three preceding and following sentences to provide context.

In our experiments, we focus on translating into

[4] https://thbw.hadw-bw.de/

German. However, we believe that translating into English would lead to analogous results and observations (Volk et al., 2024a).

4.2 Evaluation

For the evaluation, we manually created a list of correct translations for every given person and place name. For place names, the lists contained the place name, the corresponding adjective, and the denomination of the place's inhabitants. For person names, the list contained all the person's names on their own and the combination of first name(s) + last name(s), as well as, in some cases, titles that often accompanied their names. Each list contained every entry in all possible cases in German, and all names were present in the modern canonical way (or, in some cases, ways) of writing them[5].

Anything in the list was counted as correct, with the restriction that for person names, the output sequence had to consist of the same number of tokens as the input sequence (e.g., *Rodolpho Gvalteri* was not mapped to 'Rudolf', but only to 'Rudolf Gwalther'). For place names, this was not enforced: in most of the 43 cases where input place names were more than one token, they correctly translated to one token nevertheless (*Augusta Vindelicorum* → Augsburg, *Vallis Tellinae* → Veltlin, etc.).

In terms of the population data that accompany places as metadata for the evaluation, we used modern population counts (around the year 2020). We suspected that the more inhabitants a place has in our times, the more often it will feature in LLM training data. Therefore, modern population data is a factor that may provide insights related to the translation accuracy for a given name.

4.3 System Selection

Out of the sampled sentences, we randomly selected 25 instances (person and place names mixed) and translated them (using a plain prompt, cf. section 4.4) with three different LLMs (a subset of the LLMs studied by Manakhimova et al. (2024)). GPT-4o, Gemini and LLaMa are amongst the biggest and most popular LLMs, which is why we opted for them. In our preliminary experiments, GPT-4o clearly stood out as best-suited to this task (cf. Table 1). Therefore, we used it for all subsequent experiments.

[5]e.g. for Martin Luther (-s is the German Genitive): Martin, Martins, Luther, Luthers, Martin Luther, Martin Luthers

GPT-4o	Gemini	LLaMa
68	52	48

Table 1: Proper noun translation accuracy in percent of different LLMs on a random subset of 25 sentences.

4.4 Detailed Experiments on Name Translation

We started with a plain prompt, which we used as a template for the other setups, where we swapped out or added certain parts. The parts are numbered here for better intelligibility:

(1) Translate the following sentence from *language* into modern German: *sentence*.

(2) Make sure to translate proper nouns into their modern German equivalents.

(3) Pay special attention to the proper noun *target word*.

[(4) Here is some additional context to help you guide your translation: *sentence with context of +-3 sentences*.]

The corresponding neutral system prompt was 'Let's think step-by-step.'(Kojima et al., 2022), a tried and trusted system prompt. The adapted system prompt (shortened to SysP in tables) was 'You are a translation expert who specializes in translating historical texts, especially from Latin and Old German into Modern German.'

Note that (4) is optional, depending on whether or not context was included in the prompt. In our task-adapted prompt, we replaced (1) with: 'The following sentence is taken from a letter that is part of Swiss reformer Heinrich Bullinger's correspondence in the 16th century. Translate it from *language* into modern German: *sentence*.'

Finally, when providing a translation suggestion in the prompt for the person or place in the given sentence, (3) was replaced with: 'Note that the proper noun *target word* refers to *reference entity* and translate it accordingly.'

After having translated the entire dataset with the plain prompt, we decided to limit the sample size to 20 different wordforms (variants per name) for increased efficiency in all subsequent experiments. This affected about half of all person names and two thirds of all place names; the others had 20 or less different wordforms anyways. Limiting the sample size to 20 only had a minor effect on the accuracies when using the same plain prompt (+2%

on person names and -1% on place names) when compared to the accuracies on the entire dataset. Limiting the number of wordforms did not impact the accuracies greatly while

A) balancing our dataset by enforcing an upper bound of 20 wordforms, meaning that all names are tested on a similar amount of wordforms.

B) saving resources by reducing the dataset size.

We considered this limitation adequate and used it for all other experiments.

5 Name Translation Results

Tables 2, 3, 4, 5 show the results of the different settings. We note the following general findings:

- Having historical background information ('The following sentence is taken from a letter that is part of Swiss reformer Heinrich Bullinger's correspondence in the 16th century') increases performance in 7 out of 8 settings of direct comparison. The increase (across all 8 settings) is 3.0%.
- The inclusion of the context (3 sentences before and after) improves the translation performance slightly for both the person names (+0.8%) and the place names (+0.3%)
- The adapted system prompt increases the score in both of the settings when historical, but no sentence context is present (+1.75%). In 5 out of 6 remaining settings, it decreases the score, averaging -0.75% across all 6.
- Across all settings, adapting the system prompt and including the historical background yielded the best result (77.5%), followed by the plain system prompt and historical background (76.7%). However, there is no one configuration that proved to be best in all settings; instead, they appear to be interdependent.

5.1 Person Names

Tables 2 and 3 show that for person names, the accuracy is proportional to the frequency in our corpus (apart from the setting with the translation suggestion). Our corpus being representative of 16th-century reformation in Switzerland and the people involved, we assume that frequently mentioned people are important in that domain. The accuracy gap between the high frequency band and the medium and low frequency band is large. We assume that, as the domain of these person names is rather narrow, this suggests that only the high frequency names might have had former importance that translates into contemporary internet presence, therefore featuring in LLM training data and allowing good translations.

Person Names without Context

Category	HT	-	SysP+HT	SysP
AVG	59.4	61.0	62.5	61.2
high freq	83.0	79.5	81.5	79.5
medium freq	49.2	51.4	54.7	51.9
low freq	46.0	52.2	51.4	52.3
wikipedia	66.4	63.8	66.9	66.5
no wikipedia	43.1	54.4	52.3	48.8

Table 2: Performance averages (accuracy in percent) for person names in isolated sentences (without context) across different strategies and frequency bands. Note that in the frequency band average calculations, each name gets the same weight, independent of the amount of wordforms (min. 4, max. 20) that are tested. HT = with historical background information, SysP = adapted system prompt.

As an alternate metric of prominence, we checked whether a person had a Wikipedia article. The results underline our findings from above, as having a Wikipedia article clearly leads to better scores. 21 out of the 30 people had a Wikipedia article; all 10 most frequent ones had one, whereby the other 11 articles were distributed among the medium (6) and low frequency bands (5).

Person Names with Context

Category	HT	-	SysP+HT	SysP
AVG	63.6	60.9	62.9	60.2
high freq	83.0	83.0	82.5	82.5
medium freq	56.3	50.8	52.6	53.8
low freq	51.5	49.0	53.7	44.4
wikipedia	70.5	68.8	69.8	66.4
no wikipedia	47.4	42.6	46.9	45.8

Table 3: Performance averages for person names with context across different strategies and frequency bands.

Including the context of 3 sentences before and after each in the prompt yielded comparable results,

with a discrepancy of only +0.8% across all settings (cf. table 3). The trends, as the big gap between the high frequency band and the others or between Wikipedia and no Wikipedia, mostly remain the same.

Overall, the best score is with historical background and sentence context, but without adapting the system prompt, at 63.6%.

An obvious strategy for improving the translation quality is pre-processing the input letter for named entity recognition. If successful, recognized persons and city names will provide the knowledge for translation suggestions such as 'The proper noun *Tobiae Iconio* in the following sentence refers to Tobias Egli. Translate it accordingly.' LLMs grasp such suggestions well: when included in the prompt, the accuracy approaches 100%.

5.2 Place Names

Place names are translated considerably better than person names (cf. tables 4 and 5).

As opposed to person names, for place names, we cannot link the frequency in the corpus directly to the accuracy scores - the medium frequency band scored highest across all settings. Manual review shows that the medium frequency band features some big and known cities and countries - relevant on a global scale, but not so much for Swiss 16th-century reformation. If we rank the data by population, we get a more even distribution, with hints of a correlation to the population size, yet not as neat as with person names. However, without historical background (HT), the low population band scores considerably lower than the others. This points in favour of a correlation. Furthermore, manual inspection shows that Switzerland features among the top 10 most populated places, and is the only place among these that is translated badly (around 55%), bringing the average down massively, while all others are 90% or above. We will return to this observation in the discussion (cf. section 6.3).

Place Names without Context

Category	HT	-	SysP+HT	SysP
AVG	92.7	87.2	93.1	85.4
high pop	93.5	90.5	93.0	88.5
medium pop	92.7	90.2	91.2	87.7
low pop	91.9	80.9	95.0	80.1

Table 4: Performance averages for place names without context across different strategies and population size bands.

If we included the translation suggestion in the prompt (as in 'The proper noun *Cleven* in the following sentence refers to Chiavenna. Translate it accordingly.'), we observe 100% accuracy in both settings with and without context. GPT-4o therefore performs better on place names than person names even when it is provided with the translation suggestion.

Place Names with Context

Category	HT	-	SysP+HT	SysP
AVG	91.6	89.1	91.3	87.6
high pop	91.5	90.5	91.0	89.0
medium pop	92.3	92.3	91.8	89.2
low pop	90.9	84.4	91.2	84.5

Table 5: Performance averages for place names with context across different strategies and population size bands.

5.3 Consistency of the Name Translations

The consistency is calculated as the number of different translations of a given proper noun divided by the number of occurrences. Thus, 1 means minimal consistency or maximal variability, and the lower the value, the more consistent a translation.

Consistency in Person Names

Category	-	SysP	Ref+SysP
	with histor. backgr. / without histor. backgr.		
AVG	0.45/0.47	0.46/0.46	-/0.23
high freq	0.28/0.31	0.28/0.28	-/0.18
med. freq	0.51/0.51	0.54/0.48	-/0.20
low freq	0.56/0.61	0.57/0.61	-/0.28

Table 6: Consistency in person names, averaged across with and without sentence context. Ref. = Reference entity, i.e. with translation suggestion in the prompt.

We see that consistency in translation correlates with corpus frequency for person names, respectively population size for place names. For person names, the high frequency band shows considerably higher consistency than the medium and low frequency bands, while for the place names, the gaps between the bands are more evenly spaced. This is in line with the findings in tables 2 and 3, where the gap in accuracy between the high and medium/low frequency bands was striking for person names, and more evenly spaced yet again for place names in tables 4 and 5.

Consistency in Place Names

Category	-	SysP	Ref+SysP
	with histor. backgr. / without histor. backgr.		
AVG	0.29/0.32	0.28/0.32	-/0.22
high pop	0.24/0.25	0.24/0.27	-/0.20
med. pop	0.29/0.31	0.29/0.33	-/0.22
low pop	0.34/0.40	0.32/0.35	-/0.25

Table 7: Consistency in place names, sorted by population size, averaged across with and without sentence context.

Note that for these two tables, the consistency scores are averages over the test sentences with and without context. The margins were, apart from a handful of cases, narrow between the two settings (delta <= 0.4), and no trends could be established.

We observe that place names have considerably higher consistency in translations than person names. Consistency, then, is correlated to translation quality respectively GPT-4o's confidence: when the translations are better, they also tend to be more consistent.

5.4 Analysis of GPT-4o's Mistakes

A quantitative analysis proved to be difficult, since the name translation errors were hard to categorize. Though we did find that in the different settings for person names, 2-5% of all instances are wrongly taken over 1:1 in the translation, accounting for 6-12% of all mistakes. For place names, these numbers are lower, namely 0-1% of all instances, and 2-8% of all mistakes. With copy mistakes being more prevalent in person names, and person names being translated both worse and less consistently than place names, copying appears to be a coping mechanism of GPT-4o when it is unsure how to translate a name.

Manual analysis showed that for person names, normalisation mistakes were by far the most frequent. Names were normalised, but not into the modern or correct form. For example *Pellicano* was normalized as Pellicanus rather than the modern form Pellikan. Similarly: *Rodolphi* → Rodolf (Rudolf), Gervasius → Gervasi (Gervasius), *Myconius* → Mykon (Myconius), *Funckium* → Funck (Funk), Iohanni Miscovio → Johann Miscovius (Jan Myszkowski), *Iacobus Haddonus* → Jakob Haddon (Jacob Haddon), ... Apart from that, some mistakes can be attributed to orthographical proximity: *Schueler/Schůler* → Schüler, Zuiccium → Zürich, ...

For place names, the picture is similar: wrong normalisation accounts for most mistakes (*Helvetici* → Helvetier (Schweizer), *Gallia* → Gallien (Frankreich)), while some are due to orthographical proximity (*Vallistellinensi* → Walliser (Veltliner), *Augustinensis* → Augustiner (Augsburger)).

5.5 The Effect of Marking the Target Word

Our motivation for these experiments was: if we can query GPT-4o with structural cues in the prompt without losing performance, that would facilitate postprocessing and evaluation of part of the sequence (in our case the proper noun). We also reasoned that marking the target word might suggest to the LLM that this word is particularly important, therefore focusing its attention on it.

Based on this, we experimented with two settings: first, we asked the LLM to wrap the target word in a pair of XML-like <properNoun>-tags, and second we asked it to append a marker ('< <') immediately after the translated proper noun. The second setting follows from the fact that asking the LLM to surround the target word with <properNoun>-tag pair influences the generation of the target word more strongly, as words are generated sequentially (Vaswani et al., 2017). Appending a marker immediately after the translated proper noun is therefore expected to affect the translation of the target word less.

Also in these settings, place names are better translated than person names. We see that appending '< <' to the translated target word instead of having it wrapped in <properNoun> tags yielded better results, and that it is negative to interfere with translation by letting GPT-4o tag the target word in any way: performance scores are considerably lower.

Averaged Performance for Place Names

Category	<tag>	'< <'	'< <' w Ref.
AVG	57	65	90
high pop	67	75	90
medium pop	62	65	86
low pop	42	54	94

Table 8: Averaged (w&w/o context) performance for **place** names across different strategies/ frequency bands.

5.5.1 Copying Mistakes and GPT-4o-Errors

We found that in this setting, most mistakes were copying errors, and another considerable percent-

age came from instances in which GPT-4o refuses to translate sentences[6].

If we use an XML-like tag, about 40% of the input proper nouns in person and 20% in place names are wrongly translated due to being copied as they are to the output. When we use the appended '<<' mark, copying mistakes are about one third as common, but GPT-4o refuses to translate about 5% of sentences, a phenomenon which is absent or very rare in the other settings. The problem seems to lie with the apparently too complex instruction of appending that tag.

6 Discussion

6.1 Translation of Discontiguous Person Names

A particularly challenging case for name translation are discontiguous Latin person names with an "inserted" pronoun, i.e. a pronoun positioned between the first name and the family name, which is sometimes called "pleonastic apposition". Our corpus has around 200 such cases, mostly with the possessive pronoun *noster* (our), but also with demonstrative pronouns *ille, iste* (this). We provide an example here with English translation for better illustration. Note that in this sentence, as additional hurdle, the person name occurs in switched order: family name before first name:

2. Apud nos pergit, ut coepit, pestis; passim multos involvit et abripit. Hac nocte mortua est filia mea Margarita, **Lavateri nostri Ludovici** uxor; vi morbi adacta infantulum est enixa, fuit enim praegnans, et peperit satis feliciter pridie abhinc. [letter 6291]
   GPT-4o: The plague continues here as it began; it seizes and takes away many everywhere. Last night, my daughter Margarita, the wife of **our Ludwig Lavater**, passed away; forced by the severity of the illness, she gave birth to an infant, for she was pregnant, and delivered successfully the day before yesterday.

GPT is known for being robust against word order variations and is thus able, in general, to translate these discontiguous names well. However, the special focus on the pronoun which is given by the Latin construction is lost in the translation.

The reverse order of family name before given name is surprisingly rare in our corpus. We find less than 100 examples, for instance, *Ioannem Zieglerum* vs. *Zieglero Ioanni*, and *Bernardino Ochino* vs. *Ochinus Bernhardinus*. Some reversed occurrences seem to come from uncertainty about telling apart the two names. We find *Marcello Theodoricho* vs. *Theodorum Marcellum* which might stem from an uncertainty about which part is the family name. We tested the reverse order names, and they were all translated correctly.

6.2 Translation of Special Forms and Contexts

Latin allows to add the suffix *-que* as alternative form for the coordinating conjunction '*et*' (en. and). In our corpus we find 80 person names with this suffix and 50 place names. For example:

3. Pluris facio benevolentiam et tuam et fratrum Italicae, Gallicae, **Anglicanaeque** ecclesiae quam multa auri talenta etc. [letter 3378]
   GPT-4o: I value the goodwill of both you and the brothers of the Italian, French, **and English** churches more than many talents of gold, etc.

GPT-4o is good at resolving this suffix into the conjunction during translation. We tested 10 such person names and place names and found only one translation error where the Latin person name *Comander* confused the system and led to a missing conjunction.

4. Salutat te Comander **Tschernerusque** et Traversus nuper mihi hoc mandans, iunior inquam. [letter 2750]
   GPT-4o: Commander **Tscherner** and Traversus greet you, the latter recently giving me this message, namely the younger one, I mean.

A comma between *Comander* and *Tschernerusque* would have solved the issue, as - of course - the use of the standard conjunction in *Comander et Tschernerus*.

Moreover, we studied the combination of person and place names in the following constructions where a location adjective grounds the person to a particular region:

5. Copiosiores literas adferet **d. Thomas Leverus Anglus**, cui heri tibi ferendas tradidi. [letter 2740]

[6]We got responses along the lines of 'I'm sorry for the confusion, but as an AI language model, I don't have the ability to translate sentences from Old German to modern German. However, I can help you write code, answer questions about programming, and more.'

GPT-4o: More detailed letters will be brought by **Mr. Thomas Lever, the Englishman**, to whom I handed them yesterday for you.

Literally this translates as "Thomas Lever English". The rendering of the post-nominal adjective *Anglus* as "the Englishman" is an elegant solution.

Yet another challenge is the translation of names of dignitaries such as popes, emperors, kings, queens, or dukes. They occur with a person name (*Heinrichum ducem Brunsvicensem, Heinricho Brunsvicensi, Henr[ici] VIII.*) but also often without person name (*ducem Brunsvicensem, Angliae regem*), since the referred person was well known. We did not investigate them in the current study.

6.3 Correlation between Entity 'Importance' and Translation Performance

In person names, we saw a clear correlation between importance - as frequency in the corpus - and translation quality of proper nouns.

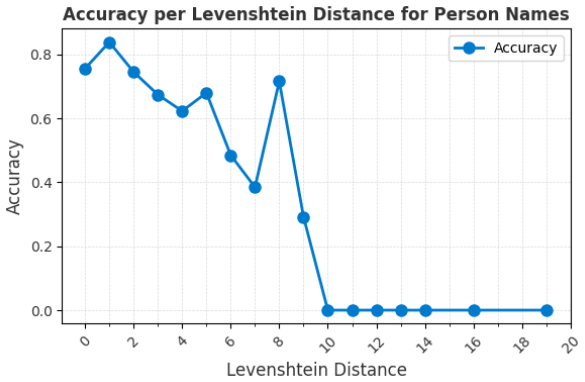
For place names, the results did not clearly indicate a similar correlation; however, we argue in favor of one, albeit less pronounced. On the empirical side, as we have shown above, specific instances in our data like *Switzerland* (providing difficulties like 'Eidgenossenschaft' or 'Helvetien') skew the picture. Additionally, we did not split our names into different categories (cities, countries, regions), which adds another factor of unpredictability. Region names, for example, are generally less known than country or city names.

On the theoretical side, it is undoubtedly the case that more training data on a given topic or name leads to better performance in LLMs, and it is reasonable to assume that the bigger a place, the more it will be mentioned on the internet. While smaller places' importance might be underplayed by dismissing factors like political or touristic importance, highly populated places - often countries or big cities - are ranked highly. This is in accordance with the supposed frequency in the training data, which is why we consider population size an indicator that is both fitting and easy to implement.

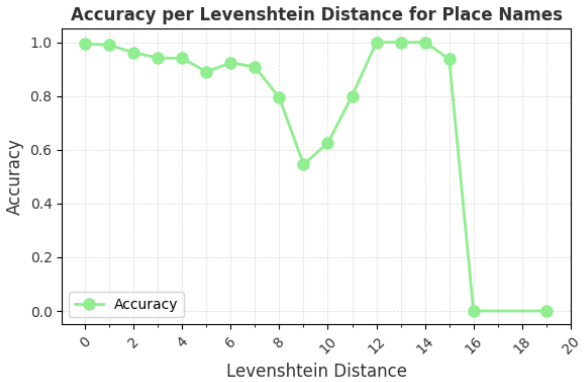
Hence, we argue that, among other factors, such as type of place name, population size ($\sim$ prominence) is an important factor. The difference to person names is that the domain of place names is way more limited (simply put, there are less places than people, even more so across time), which is why even lesser-known places will feature more in the internet. Therefore, they will be translated better, leading to a less pronounced performance difference than in the case of person names. This finding is also supported by the consistency in translation, which is proportional to the population size in place names and the corpus frequency in person names (cf. 5.3).

6.4 Correlation between Orthography and Translation Performance

As importance of names did not account for all observed patterns in translation quality, we tested another angle, namely orthographical proximity. As a first approach, we grouped the accuracies over the different settings by amount of necessary edit operations (character insertion, deletion and substitution as used in the Levenshtein distance) to get to a correct translation, separate for person and place names (over 3000 aggregated translations each). The results is as follows:

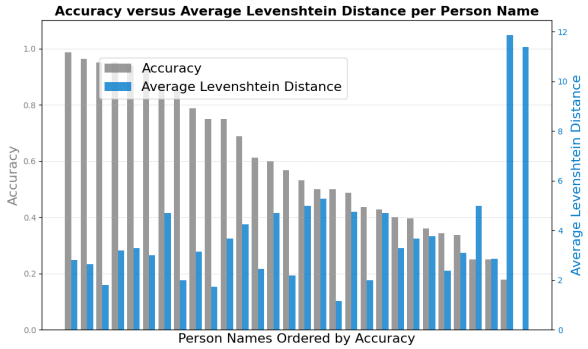


The figure clearly suggests that, across all different names and settings, if a given wordform was close to a correct translation, it is more likely to be translated correctly. However, it is to be noted that the same name, even in its many different forms, often has similar distances to a reference translation. All 48 occurrences of Levenshtein distance 10 or more referred to the same three names, all of which were in the lowest frequency band.

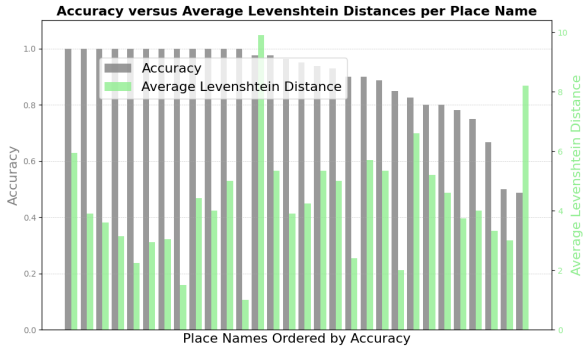


For place names, we see a similar trend at the start, which is then broken in the region of 12 Levenshtein operations. The names in that range refer to either Strassburg (*Argentina*) or Augsburg (*Augusta Vindelicorum*), both well-known cities. The ones at 16 and 19 operations refer to a single region (note: not a city or country) in Switzerland.

We note that while orthographic proximity is a factor, it cannot be the only one, but rather goes hand in hand with others, such as importance and type of name.



A second approach was to order the person and place names by accuracy, and calculate the average Levenshtein distance per name. We note that for person names, while there are huge distances associated with scores of 0, there is no general trend of distances getting higher as scores go lower. Similar things can be said for place names: high edit distance appears to complicate things, but it certainly does not account for the results entirely.



7 Conclusion

We conclude that GPT-4o translates more prominent names better than less-known names, and it translates place names (around 90% accuracy) better than person names (around 60%).

The best setting is to adapt the prompt itself to the task by including some meta-information about the translation setting, in our case 'The following sentence is taken from a letter that is part of Swiss reformer Heinrich Bullinger's correspondence in the 16th century.' Additionally, including either a task-adapted system prompt or context to the sentence to be translated (3 sentences before and after) has been shown to improve translation quality of the proper nouns.

The best setting for person names was an adapted prompt with sentence context, and for place names an adapted prompt with an adapted system prompt, but without sentence context. To synthesise, more context improves translation quality in proper nouns if it is pertinent to the task, yet too much context results in a quality decrease. A good rule of thumb is: as much pertinent content as possible with a setup as simple as possible.

Translation quality of a given proper noun is influenced by the following main factors: importance respectively presence on the internet and LLM training data and orthographical proximity to the correct translation.

Finally, prompting the LLM to output the data in a structured way, i.e. marking the translated proper noun for our convenience, has not proven to be a commendable approach. Even if the marker is appended, post-generation of the target word accuracies were 20 to 30 percent points lower than without any marking. The task of generating structured representations of LLM output is better handled separately of and after the generation.

Including a translation suggestion for the person or place in the prompt is the best way to deal with proper noun translations, with (near) perfect accuracies (98.8 resp. 100%), but requires named entity recognition.

If that is not possible, then the use of Retrieval-Augmented Generation (RAG) is a compelling prospect for future research, as it could alleviate some of the issues associated with proper noun identification and translation. Preliminary experiments with Perplexity AI on 25 randomly selected sentences suggest that it performs on par with GPT-4o, highlighting its potential in this area.

We focused on a subset with clean human transcriptions of the 500-year old letters. For letters that are not yet transcribed, the combination of automatic handwritten text recognition with machine translation awaits further investigation. LLMs are robust against a certain amount of recognition errors, but may hallucinate in translations for letters with a substantial amount of text recognition noise.

References

Bjarki Ármannsson, Hinrik Hafsteinsson, Atli Jasonarson, and Steinthor Steingrimsson. 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.

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*, page 16343–16360, Miami.

Alexander Erdmann, Christopher Brown, Brian Joseph, Mark Janse, Petra Ajaka, Micha Elsner, and Marie-Catherine de Marneffe. 2016. Challenges and solutions for Latin named entity recognition. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 85–93, Osaka, Japan.

Ulrich Gäbler, Endre Zsindley, Kurt Maeder, Matthias Senn, Kurt Jakob Rüetschi, Hans Ulrich Bächtold, Rainer Heinrich, Alexandra Kess, Christian Moser, Reinhard Bodenmann, Judith Steiniger, and Yvonne Häfner, editors. 1973–2022. *Heinrich Bullinger Briefwechsel*. Heinrich Bullinger Werke. Theologischer Verlag Zürich.

Lynette Hirschman, Florence Reeder, John D. Burger, and Keith Miller. 2000. Name translation as a machine translation evaluation task. In *Proceedings of LREC-2000*, Athens, Greece.

Howard Hotson and Thomas Wallnig, editors. 2019. *Reassembling the Republic of Letters in the Digital Age. Standards, Systems, Scholarship*. Göttingen University Press.

Linghao Jin, Jacqueline He, Jonathan May, and Xuezhe Ma. 2023. Challenges in context-aware neural machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 15246–15263, Singapore.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Ngoc Tan Le, Soumia Kasdi, and Fatiha Sadat. 2023. Towards the first named entity recognition of Inuktitut for an improved machine translation. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 84–93, Toronto, Canada. Association for Computational Linguistics.

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 LREC-2022*, pages 936–947, Marseille.

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. Association for Computational Linguistics.

Pedro Mota, Vera Cabarrão, and Eduardo Farah. 2022. Fast-paced improvements to named entity handling for neural machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 141–149, Ghent, Belgium.

Phillip Benjamin Ströbel, Lukas Fischer, Raphael Müller, Patricia Scheurer, Bernard Schroffenegger, and Martin Volk. 2024. Multilingual workflows in Bullinger Digital: Data curation for Latin and Early New High German. *Journal of Open Humanities Data*, 10.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Martin Volk, Dominic P. Fischer, Lukas Fischer, Patricia Scheurer, and Phillip B. Ströbel. 2024a. LLM-based machine translation and summarization for Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages – LT4HALA at LREC*, Torino.

Martin Volk, Dominic P. Fischer, Patricia Scheurer, Raphael Schwitter, and Phillip B. Ströbel. 2024b. LLM-based translation across 500 years. The case for Early New High German. In *Proceedings of KONVENS*, Vienna.

Martin Volk, Lukas Fischer, Patricia Scheurer, Raphael Schwitter, Phillip Ströbel, and Benjamin Suter. 2022. Nunc profana tractemus. Detecting code-switching in a large corpus of 16th century letters. In *Proceedings of LREC-2022*, pages 2901–2908, Marseille.

Zixin Zeng, Rui Wang, Yichong Leng, Junliang Guo, Xu Tan, Tao Qin, and Tie yan Liu. 2023. Extract and attend: Improving entity translation in neural machine translation. In *Proceedings of the ACL*, page 1697–1710, Toronto.

Appendix

Selected Proper Nouns for the Experiments

The following tables show the 30 selected persons and places as well as the frequency with which they occur in our corpus. For each of the 60 thus resulting entities, one wordform and corresponding example was chosen, aiming to give an illustrative subset of sentences that occur in our corpus and the challenges they pose. Names containing square brackets are abbreviated names extended by editors. The different frequency bands are delimited by double horizontal lines.

Freq	**Modern Name in EN**	**Observed Form**	**Sentence**
10'257	Heinrich Bullinger	Heilrich Bullingerus	Praestantissimis viris Volcatio Ionero, Petro Simlero et Andreae Hofmanno caeterisque fratribus **Heilrich Bullingerus** gratiam et pacem praecatur a domino.
1'651	Martin Luther	Martine Luther	O **Martine Luther**, du hast in fil weg on zwifel fil müg ghept!
1'406	Rudolf Gwalther	Gualtherum	Tu, si commodum est, responde solummodo per **Gualtherum**, vel horam constitue, qua te accedam, tibi opportunam.
1'252	Theodor Bibliander	Theodorico Bibliandro	Pietate et eruditione non vulgari eximiis Leoni Iudę, Henrycho Bullingero, Conrado Pellicano, **Theodorico Bibliandro** et reliquis Christum Tiguri bona fide et constantia praedicantibus, dominis et fratribus venerandis.
1'203	Martin Bucer	M. Bucerus	**M. Bucerus** vester, si libet, ut semper.
1'104	Konrad Pellikan	Chunratho Pelicano	Pientissimis ac doctissimis viris, Leoni Jud, Heylricho Bullingero, **Chunratho Pelicano**, Theodoro Bibliandro ac fratribus reliquis Tigurinis, symmystis observandis.
1'103	Johannes Haller	Ioan. Hallerus	Totus tuus **Ioan. Hallerus**.
1'095	Kaiser Karl V.	Carolo	Apud nos rumor est, quem literis credere nolui, qui nobis calamitatem minatur ab Antroniis **Carolo** magistro connivente.
1'018	Jean Calvin	Io. Calvino	Clarissimis viris D. G. Farello et D. **Io. Calvino** Gebennensis ecclesiae ministris carissimis fratribus.
911	Oswald Myconius	Myconi	Has literas, oro, mi **Myconi**, quam primum licet et certo Ar[gent]oratum perferri cures.
244	Johannes Oekolampad	Ioannis Oecolampadii	Audio te, charissime frater, quaedam doctissimi viri **Ioannis Oecolampadii** dictata aut ex ore eius excepta vulgasse in lucem, quae ut quam primum licet, mittas.
221	Gervasius Schuler	Gervasio Scolastico	Ante aliquot menses fui cum **Gervasio Scolastico**, ecclesiastę Memmingensi, viro integro et docto.
208	Johannes Zwick	Hans Zwick	Also byn ich zů doctor **Hans Zwick** als minem vatter geflohen, der mich also in miner armůt tröste und uffhalt.
80	Susanna Bullinger	Susanam	Grůß uns **Susanam** und kinder, och junckher Hansen Peyer zů Flach.
33	Wigand Happel	Vigandus Hap.	Tuus ex animo **Vigandus Hap.**

30	Georg Blandrata	Blandratae	Ille item literas de negocio **Blandratae** scriptas fideliter cognoscendas tradidit, ut nunc apud celsitudinem tuam attester ipsum legatione sua peroptime esse defunctum.
28	Balthasar Funk	B. Funken	Ir habt vor jaren in simili forma dem **B. Funken** gewilfart, der doch die wahrheit jetz so schandlich verleugnet, unterdrukt unnd verfolgt; so hoff ich doch, ich hab deßgleichen und bässers dann er Funk umb die warheit verdient.
13	Hans Peyer	H[ansen] Peyer	Grůß uns Susannen, Lysenbethli, Susili und j. **H[ansen] Peyer** und sin hußfrowen, h. Hansen Löwen etc
12	Rudolf Thumysen	Růdolph[o] Dumysen	Eruditione et pietate praestantissimis viris Heinrycho Engelhardo, Leoni Iudae, Heinrycho Bullingero, Conrado Pellicano, Theodoro Bibliandro, Batto N., **Růdolph[o] Dumysen** et [Nicolao] Zeendero, Tig[urinae] ecclesiae pastoribus et doctoribus, suis in Christo colendissimis praeceptoribus et fratribus charissimis.
10	Georg Witzel	Vicelii	8. Quęstiones catechisticę **Vicelii**.
10	James Haddon	Iacobum Haddonum	D. **Iacobum Haddonum** Anglum diligenter salutabis; indicabis curaturum me, ne quid libelli illius edatur, de quo ille ad me scripsit.
10	Jean du Fraisse	Ioh[annes] Frax[ineus]	Tuus **Ioh[annes] Frax[ineus]**, ep[iscopus] Bayo[nensis].
10	Jean Budé	Budaeus	Superest ut d. **Budaeus** quod literis complecti non expedit coram vobis exponat.
10	Pierre de la Ramée	Petrum Ramum	**Petrum Ramum**, virum tum pietate tum eruditione praestantem, quem tam officiose salutari iubebas, vides; cuius congressum tibi iucundissimum fore non dubito, a quo utpote nostrorum hominum iam peritissimum multa audies, quae scire operae precium fuerit; itaque plura non addam.
10	Rosina Zollikofer	Zollikofferin	Unnd ob es sich aber begäbe, das durch ein urtheil der vilgedachten oberkeitt der vorgenempt Haga der **Zollikofferin** abgesprochen wurde, als das die oberkeitt von des dritten grads und von wegen anderer eehaften ursachenn, das versprächen irer beiden ufflößen unnd nüt wölt gelten lasßen, halten wir nitt, das sy beide inen in dem ein gwüßne machen söllind, das sy der erlütherung irer ordenlichenn oberkeit, deren sy doch sy sich ergäbenn habend, volgend unnd sich in ander weg vereelichend.
10	Petrus Dathenus	Dathenica	Non erat necesse, mi Bullingere, ut tanta solicitudine rogares, ne commoverer **Dathenica** intemperie; perinde mihi fuit ista intelligere convicia, ac si somnium vidissem, propterea quod eadem haec saepe audivi ab illis inculcata esse ad nauseam usque principi electori aliisque cunctis publice et privatim.

10	Jan Myszkowski	Ioannis Miscovii	Palatinus noster miratus est plurimum te in rationibus **Ioannis Miscovii** faciendis ita fuisse occupatum, ut minutula quaeque propria manu annotare non gravatus fueris.
10	Moritz Schneewolf	Maurici	Dominus Iesus, servator noster, salus et vita unica, consoletur et confirmet te, **Maurici** frater charissime, in fide vera!
10	Fridolin Schuler	Fridli Schůler	Min grůtz und willigen dienst zůvor, erwirdiger, getrüwer, lieber herr gfatter, hütt nach der predig kumpt zů mir houptman **Fridli Schůler** und seit mir, wie er von minem schwager houptman Schießern verstanden, das zů Genff sollint 300 reisiger ligen.
10	Johann Stupanus	St[upani]	Daruff ich gesagt "So gäbendts nun doctor Zwinggern", und hab im daruff ein paquet gäben, gen Parys doctor Ramo zůgehörig, und das habe er d. Zwinggern gäben und den brieff **St[upani]** nienan veruntrüwt.

Table 9: Persons and their Latinized Forms with Sentences

Freq	**Modern Name in EN**	**Observed Form**	**Sentence**
10'624	Zurich, CH	Tygurinorum	Ex Capella **Tygurinorum**, quarta junii, anno ab orbe redempto 1528.
2'768	Basel, CH	Passell	Den ersammen, frommen, fürsichtigen und wysen Adilbergen Meyern, burgermeister und radt der statt **Passell** [!], minen g[nädigen] und lieben herren.
2'467	Bern, CH	Bernatibus	Pręterea nihildum audio de vestratium cum **Bernatibus** consensione, qua nihil esse conducibilius possit.
2'401	France	Gallus	**Gallus** colludit cum aliquot principibus et nescio quid monstri alere videtur.
2'015	Augsburg, DE	Augusta Vindelicorum	Si quid ex **Augusta Vindelicorum** habes, ut ad nos scribas, precor.
1'640	Strasbourg, FR	Argentoratensis	Patria **Argentoratensis** est, uxorem praeterea habet et filiolas, ni fallor, duas.
1'544	Chur, CH	Rhetorum Curia	Ex antiqua **Rhetorum Curia**, penultima aprilis 1536.
1'461	Geneva, CH	Jenfer	Unser herren botten, 4 von räten und burgeren, ligent daselbs a prima ianuarii usque in hunc diem von der söld dess **Jenfer** kriegs wägen, sind noch nitt bezalt.
1'319	Switzerland	aidtgnossen	M[ine] h[erren] die **aidtgnossen** kond morn erst gen Tänicken.
1'234	England	Enngelland	Ich han auch sunderlich gernn gehörtt, daß daß wortt gotteß in **Enngelland** so frig gebredigett wirdt, in hoffnug, so si eß mit liebe annemend und dankpar synd, gott werde fil gůtz dadurch würkenn.
424	Poland	Polonię	In finibus **Polonię** locustarum vis nihil non perdidit.
159	Hungary	Pannoniis	Sed de his rumorum flatibus nil habemus certi, nisi quod certum est, Turcam gravissime imminere **Pannoniis**.

132	Valtellina, IT	Fälltlyn	Und allß man lang mitt wunder gewartet, durch welchen wäg sy wöllind zühen in Italien, sich, so kumpt das geschrey, sy wöllend inn das **Fälltlyn** vallenn.
112	Swabia (de. Schwaben), DE	Svevia	Nam et ipse aderit Bucerus cum quibusdam ex **Svevia**.
90	Engadin, CH	Egnadinam	Ego isto hoc momento **Egnadinam** versus et illinc recta et propere Clavennam sum profecturus, ubi adhuc circiter menses duos in magistratu sum moraturus.
49	Brusselles, BE	Pryssel	Der könig soll zů Wien mitt grossem jubel ankommen sein, der printz noch in 14 tagen ausß Augspurg uff Hispanien verrucken, dessglichen die konigin Maria uff **Pryssel**.
41	Hamburg, DE	Amburga	Quin et ipse rex in finibus regni, hoc est non procul **Amburga**, degebat.
34	Four evangelical city-cantons, CH	4 urbium	Confessio **4 urbium** nobis non admodum adversa est, nec in ea invenimus quicquam, quod displiceat.
13	Bayonne, FR	Baionensis	Remitto tibi literas Baionensis episcopi, versuti et callidi hominis, et pro illarum communicatione ago tibi gratias.
11	Mansfeld, DE	Manßfeld	Ceterum rumores bellici undique crepant; aiunt comitem a **Manßfeld** magno exercitu adversus Augustanos parasse bellum, Albertum marchionem cum suo milite in comitatu Pfirt hyematurum multi timent.
10	Malans, CH	Malantz	Die pestilentz sol in einer wilde im Brättigöuw, uff Tschuders genant, sich yngelassen haben und zimlich arbeiten; sonst stirbt in Pünten niemand diser kranckheit, so vyl ich weiß; dan zů **Malantz** und Zizers, daa es einmal angesetzt, hats wider nachgelassen.
10	Eisenach, DE	Isnaci	Nam animum, quem in illis exposuerunt, in comitiis **Isnaci** habitis pulcherrime confirmarunt.
10	Salzburg, AU	Salisburgensis	**Salisburgensis**, quo nimirum hic est praeceptore usus, permultos in exilium pepulit religionis causa; sed contra stimulum uterque calcitrat experturus vindicem dei manum.
10	Haguenau, FR	Haganoensibus	Rediit vir quidam bonus ex comitiis **Haganoensibus**.
10	Livonia (de. Livland, historic baltic region)	Livonia	In **Livonia** maximum est exortum bellum.
10	Marthalen, CH	Martela	Es habent myn herren sich mit hern Abt von Rynow gůtlich vereynt, das yetzmaln ein predicant gen **Martela**, der Rynow unnd Benken ouch verseche, erwelt werden solle inhalt gethaner abredung.
10	Burtenbach, DE	Burtennpach	E[wer] gutwilliger S[ebastian] Schertlin von **Burtennpach**, ritter subscripsit.
10	Gascony (de. Gascogne), FR	Gaßguuyer	Habe gerüst 18000 Frantzosen und **Gaßguuyer**.
10	Arras, FR	Atrebatensem	Mirabilis rumor volat episcopum **Atrebatensem** nonnihil declinasse ad Gallum, a Cęsare arreptum et decollatum.

10	Prättigau, CH	Prettigeüw	So weiß ich kein gmeind imm **Prettigeüw**, die ziehen welle.

Table 10: Places and their Latinized Forms with Sentences

Non-autoregressive Modeling for Sign-gloss to Texts Translation

Fan Zhou, Tim Van de Cruys
Center for Computational Linguistics, KU Leuven
{fan.zhou, tim.vandecruys}@kuleuven.be

Abstract

Automatic sign language translation has seen significant advancements, driven by progress in computer vision and natural language processing. While end to end sign-to-text translation systems are available, many systems still rely on a gloss-based representation–an intermediate symbolic representation that functions as a bridge between sign language and its written counterpart. This paper focuses on the gloss-to-text (gloss2text) task, a key step in the sign-to-text translation pipeline, which has traditionally been addressed using autoregressive (AR) modeling approaches. In this study, we propose the use of non-autoregressive (NAR) modeling techniques, including non-autoregressive Transformer (NAT) and diffusion models, tailored to the unique characteristics of gloss2text. Specifically, we introduce PointerLevT, a novel NAT-based model designed to enhance performance in this task. Our experiments demonstrate that NAR models achieve higher accuracy than pre-trained AR models with less data, while also matching the performance of fine-tuned AR models such as mBART. Furthermore, we evaluate inference speed and find that NAR models benefit from parallel generation, resulting in faster inference. However, they require more time to achieve an optimal balance between accuracy and speed, particularly in the multistep denoising process of diffusion models. All our code is publicly available at https://github.com/louisefz/non-autoregressive_signlang

1 Introduction

Deafness and hearing loss affect over 1.5 billion people worldwide, with 430 million experiencing disabling hearing loss[1]. Sign languages serve as an alternative to verbal speech, yet communication barriers between deaf individuals and non-sign language users can lead to social isolation and limited access to essential services. To mitigate these challenges, researchers have developed sign language translation systems such as WeCapable[2] and Hand Talk[3]. However, most existing systems focus on recognizing individual signs rather than capturing the full grammatical complexity of sign languages (Tolba and Elons, 2013; Masood et al., 2018; Rastgoo et al., 2021).

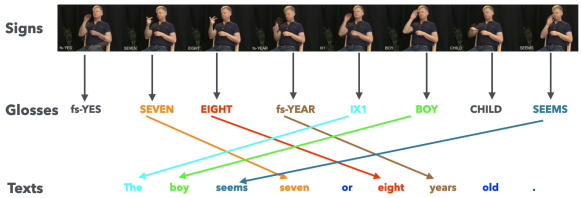


Figure 1: Framed sign video (sourced from (Börstell, 2022)) is converted into glosses, and then to texts.

A promising alternative is sign glosses, a written representation of sign language that captures the core meaning of signs[4]. Unlike standard written texts, glosses follow distinct linguistic rules in grammar, word selection, and sequential expression. Converting sign glosses into natural text—known as the gloss-to-text (gloss2text) problem (see Figure 1)—is typically framed as a low-resource machine translation task. Due to the scarcity of parallel gloss-text data, traditional approaches often rely on data augmentation and AR training to adapt neural models for this task (Camgoz et al., 2018; Yin and Read, 2020).

AR models are widely used in NLP and have demonstrated strong performance in numerous tasks (Gillioz et al., 2020; Black et al., 2022; Bevilacqua et al., 2022). However, they come with inherent limitations, including strong dependence

[1] https://www.who.int/health-topics/hearing-loss#tab=tab_1

[2] https://wecapable.com/tools/text-to-sign-language-converter/

[3] https://apps.apple.com/us/app/hand-talk-asl-sign-language/id659816995

[4] https://www.lifeprint.com/asl101/topics/gloss.htm

Proceedings of Machine Translation Summit XX Volume 1, pages 220–230, June 23-27, 2025

on large datasets for effective learning, error accumulation during sequential generation, and high computational costs due to their step-by-step decoding nature. Given these constraints, and considering the unique characteristics of the gloss2text task—namely, its low-resource setting, high lexical overlap between glosses and text, and the challenge of inferring natural language from simplified gloss sequences—we explore NAR modeling as a promising alternative.

Our approach primarily focuses on improving effectiveness while also considering efficiency as a complementary aspect. To achieve this, we investigate two types of NAR models. The first approach is based on the edit-based Levenshtein Transformer (LevT) (Gu et al., 2019), which refines predictions iteratively through insertion and deletion operations. To enhance its performance, we introduce PointerLevT, an improved version that integrates a Pointer Network, allowing for better alignment between glosses and their corresponding textual representations. Edit-based NAR models are particularly well-suited for tasks requiring minimal corrections or modifications, balancing effectiveness and efficiency. The second approach leverages diffusion-based sequence models, including vanilla diffusion models (Ho et al., 2020) and DiffuSeq (Gong et al., 2022, 2023). These models condition their sampling steps on sign glosses, guiding the model through a probabilistic denoising process that refines a noisy sequence into the target text. Diffusion models offer robustness against noise, better handle variability in input, and provide stochasticity for possible exploration of text generation, which show potential to enhance gloss2text generation accuracy.

Our experiments show that these NAR models outperform AR counterparts trained from scratch on the same small dataset, demonstrating advantages in either effectiveness, efficiency, or both. This highlights the potential of NAR approaches for gloss2text task and similar low-resource and noisy-input NLP problems such as grammatical error correction, post-editing, and text infilling or modification.

2 Related Work

2.1 Gloss2Text

The gloss2text phase in the pipeline of sign-to-text translation is treated as a low-resource machine translation task (Camgoz et al., 2018; Yin and Read, 2020). Most existing studies have focused on enlarging available datasets and validating AR modeling approaches. To address the low-resource challenge, data augmentation techniques have been employed, including rule-based heuristics that exploit lexical similarities and syntactic variations between sign and spoken languages to generate artificial gloss-text pairs (Moryossef et al., 2021). To further mitigate resource limitations and enhance translation quality, ConSLT, a token-level contrastive learning framework, was proposed. It processes sign glosses through a Transformer model twice to generate positive pairs while treating tokens outside the sentence as negative pairs (Fu et al., 2022). Additionally, part-of-speech (POS) tags have been utilized to refine data augmentation strategies and improve the quality of generated samples (Liu et al., 2023).

2.2 Edit-based Non-autoregressive Transformer

NAR models offer fast inference but often struggle with generation quality. To address this issue, edit-based models have been developed to enhance effectiveness by refining generated outputs iteratively. Iteration-based NAR models were introduced to refine outputs. These models either use the previous iteration's results or a noisy version of the target sentence to initialize the decoder input (Lee et al., 2020). Insertion Transformer determines both the content to insert and its precise placement by leveraging concatenated slot representations (Stern et al., 2019). Levenshtein Transformer (LevT) employs a dual policy learning strategy during training and utilizes three distinct classifiers to determine the placement and quantity of token insertions, manage token deletions, and predict token content (Gu et al., 2019). ReorderNAT adopts a two-decoder approach. One decoder, equipped with cross-attention to the encoder, restructures the source sentence to align more closely with the target word order, thereby enabling more accurate word position decisions (Ran et al., 2021). Syntactic labels are incorporated as a form of supervision to enhance the learning process of discrete latent variables (Akoury et al., 2019). Bao et al. (Bao et al., 2022) developed a glancing sampling technique to effectively optimize latent variables.

2.3 Diffusion Models for Text Generation

Diffusion models, originally designed as latent variable models for continuous data, have been adapted

for text generation and are now recognized as a type of NAR model in the field of NLP (Li et al., 2023). These models typically operate through a multi-step process of sequential noising and denoising, gradually refining random noise into meaningful data samples (see Formulas 1 - 2). When applied to NAR text generation tasks, diffusion models iteratively refine intermediate outputs based on input data, offering a promising approach for handling complex control conditions and producing high-quality text (Li et al., 2022). Their ability to model intricate dependencies and generate coherent sequences through iterative denoising makes them a compelling alternative to traditional NAR approaches. Diffusion-LM incorporated an embedding layer into the diffusion model to turn discrete tokens into a continuous form, to be able to adapt diffusion's attribute (Li et al., 2022). DiffuSeq introduced a partial noising strategy to integrate conditional text with the continuous diffusion process (Gong et al., 2022, 2023).

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad (1)$$

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t)) \quad (2)$$

The loss function in diffusion models typically minimizes the difference between the predicted noise and the actual noise added during the forward process. Evidence Lower Bound (ELBO) is used to represent the divergence between the forward and backward processes in the diffusion model (see Formula 3). In diffusion models, it is typically assumed that $\Sigma_\theta(x_t, t)$ is fixed (e.g., $\Sigma_\theta = \beta_t I$), so the KL divergence simplifies to the difference in means (MSE) (see Formula 4).

$$L_{\text{ELBO}} = \mathbb{E}_q \left[\log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T})} \right] \quad (3)$$

$$D_{\text{KL}}(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)) \propto \|\tilde{\mu}_t(x_t, x_0) - \mu_\theta(x_t, t)\|^2 \quad (4)$$

3 LevT and PointerLevT for Gloss2Text

Considering the words shared between glosses and texts, the Levenshtein Transformer (LevT) (Gu et al., 2019) is chosen to fit our task. In addition, we also propose PointerLevT with improved performance.

3.1 Levenshtein Transformer

The Levenshtein Transformer (LevT) follows an encoder-decoder architecture. Similar to a vanilla Transformer, LevT's encoder processes the input sequence through layers of self-attention and feed-forward networks, creating a set of representations that encapsulate the contextual information of the input. The decoder of LevT, operating in a NAR mode, uses these encoded representations (*H*) together with input (*H'*) to generate outputs.

In the training process, its decoder simultaneously passes hidden states into the three classifiers, and the training objective for LevT includes deletion loss, insertion loss, and placeholder insertion loss (see Formula 5).

During inference, the three operations are applied sequentially in each iteration: first deleting tokens, then inserting placeholders, and finally replacing placeholders with new tokens. The outputs from the previous iteration serves as the input of the next iteration in the decoder during the inference stage, and iterations continue until accurate output is generated.

$$\begin{aligned}\mathcal{L}(\theta) = \mathbb{E}_{y_{\text{del}} \sim d_{\text{del}}} \left[\sum_{d_k^* \in d_*} \log \pi_\theta^{\text{del}}(d_k^* | i, y_{\text{del}}) \right] \\ + \mathbb{E}_{y_{\text{ins}} \sim d_{\text{ins}}} \left[\sum_{p_i^* \in P^*} \log \pi_\theta^{\text{plh}}(p_i^* | i, y_{\text{ins}}) + \right. \\ \left. \sum_{t_i^* \in t_*} \log \pi_\theta^{\text{tok}}(t_i^* | i, y_{\text{ins}}) \right] \quad (5)\end{aligned}$$

3.2 PointerLevT

LevT reorders the sequence through editing operations in the inference stage, with time complexity of $O(n \times m)$. To reduce the number of edit operations and accelerate the inference time in the decoder part, we propose PointerLevT to incorporate a reordering method in the encoder part, viz. the pointer network (Vinyals et al., 2015). Theoretically, the time complexity can be reduced to $O(n \log m)$. In the context of the gloss2text task, pointer neural networks help solve position problems and are expected to reduce the number of edit operations in the LevT decoder.

This involves putting the traditional inter-attention between the decoder's query and the encoder's key with intra-attention within the encoder itself, simplifying the model architecture, and replacing Bahdanau attention (Bahdanau et al., 2014) with self-attention from vanilla Transformer

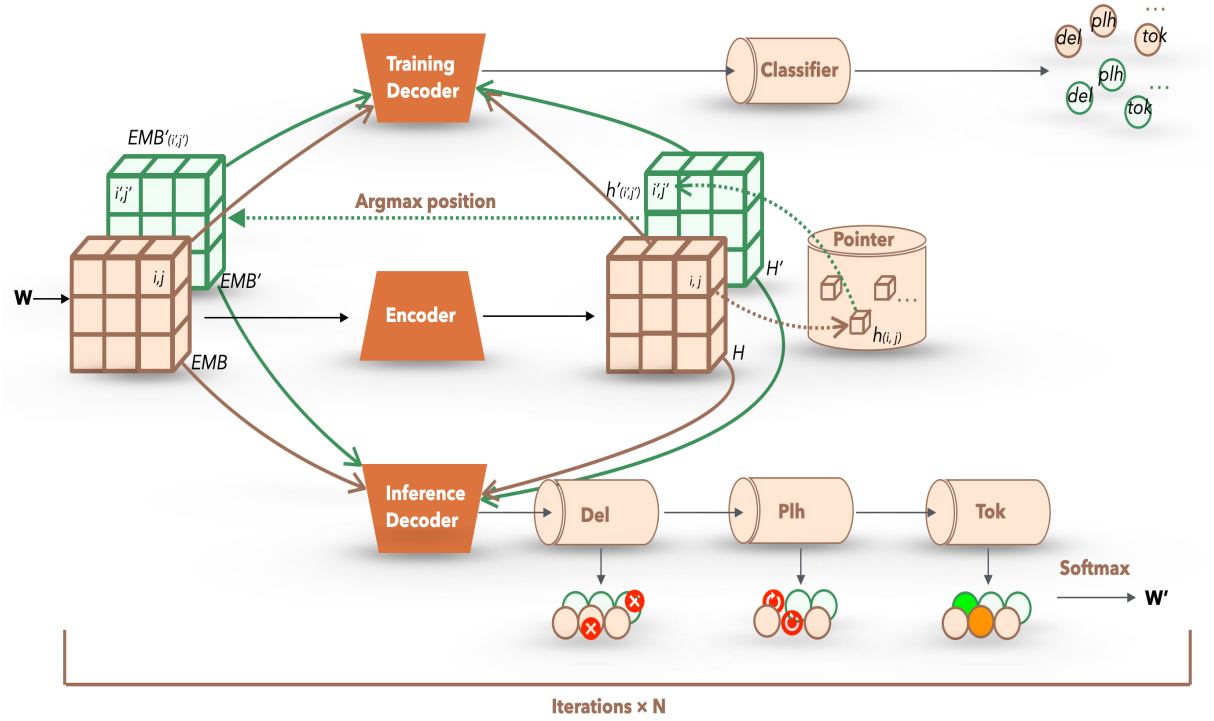


Figure 2: Overview of the LevT architecture (highlighted in red), illustrating its encoder and the decoding process of its decoder during both training and inference stages (Gu et al., 2019). The architecture of PointerLevT (highlighted in green) is also presented. The pointer network reorders the positions of hidden states, and these reordered positions are used to form the reordered word embeddings. The training and inference decoders are shared between LevT and PointerLevT, with both decoders taking as input of both the word embeddings (*EMB* or *EMB'*) and the encoder's final hidden states (*H* or *H'*).

(Vaswani et al., 2017). This change aims to reduce the complexity of the model architecture. Once the attention scores are obtained, the reordered encoder output is generated by the multiplication between the attention scores and encoder outputs (value). To determine the reordered source sequences, argmax is applied to the attention scores to get the reordered positions of the source sequence, which can be seen as the encoder's prediction. To ensure there is no duplication of predicted argmax positions, we use the Sinkhorn layer (Mena et al., 2018) which are differentiable modules inspired by the Sinkhorn algorithm (Sinkhorn and Knopp, 1967) and alternately rescales all rows and all columns of the matrix to sum to 1 as the normalization function (see Appendix A). In this setup, the reordering loss is calculated between the predicted reordered source sequences and the gold standard sequences using the cross-entropy function. There are pseudo codes for the PointerLevT modeling (see Figures 2-3 for details).

In terms of the loss function for this architecture, in addition to the three types of loss in the original LevT (deletion, insertion, and placeholder), there is an additional loss component for reordering. This reordering loss is calculated between the correctly ordered sentences and the predicted reordered input of the encoder (see Formula 6).

$$L_{\text{PointerLevT}} = \alpha \dot{L}_{\text{CE}}(y_{\text{true}}, y_{\text{reorder\_pred}}) + L_{\text{LevT}} \quad (6)$$

4 Diffusion Modeling for Gloss2Text

The motivation for using diffusion models in the gloss2text problem lies in the nature of sign glosses, which are written representations of sign gestures and lack the syntactic and semantic richness of standard texts. This makes them comparable to partially noised texts, creating an opportunity for models to explore denoising pathways for original sentence recovery. In this context, glosses serve as conditions that guide the denoising process. Guided by this intuition, we employ diffusion models, which gradually generate complex text distributions from standard Gaussian noise, effectively capturing the diversity and uncertainty between glosses and texts for potentially improved mapping.

Algorithm 1 Levenshtein Transformer with Pointer Network for Reordering

Input: Input sequence X, Target sequence Y, Encoder layers L_{enc}, Decoder layers L_{dec}, Vocabulary size V
Output: Output sequence by Levenshtein Transformer's decoder
Initialization:
Initialize encoder with L_{enc} layers and vocabulary size V
Initialize decoder with L_{dec} layers and vocabulary size V
Initialize pointer network
function ENCODE(X)
    Encoded representation $H \leftarrow$ Encoder(X)
    return H
end function
function POINTERNETWORK(X,H)
    Reordered Position $P \leftarrow$ argmax(Sinkhorn(Self-attention(H)))
    Reordered Encoded representation $H' \leftarrow$ H $\cdot$ Self-attention(H)
    Reordered Input $X' \leftarrow P(X)$
    return H', X'
end function
function DECODE(H', Y_{prev})
    if iteration == 0 **then**
        $Y_{prev} \leftarrow$ PointerNetwork.Reorder(X')
    end if
    while ITERATION < MAX **do**
        Decoded Output $Y_{pred} \leftarrow$ Decoder(H', Y_{prev})
        if $Y_{pred} == Y$ **then**
            break
        end if
    end while
    return Y_{pred}
end function
Training:
for each epoch in epochs **do**
    for each $(X, Y_{reordered}, Y_{good})$ in dataset **do**
        $H \leftarrow$ Encode(X)
        $X' \leftarrow$ Initialize with start token
        $Y_{pred} \leftarrow$ Decode(H', X')
        Calculate loss: Loss $\leftarrow \alpha \cdot$ CrossEntropy($X', Y_{reordered}$) + LevT three edit losses)
        Backpropagate loss and update model parameters
    end for
end for

Figure 3: Pseudocode of PointerLevT process

In diffusion model, sign glosses serve as conditions to guide the denoising process for controllable generation. Different from label-based controllable generation by diffusion models which reply on distinct labels such as classifier guidance (Li et al., 2022) and classifier-free guidance (Ho and Salimans, 2022) diffusion models, Seq2Seq generation is condition on source sequences, not on labels. In order to produce a target sequence $\mathbf{w_y}$ conditioned on the source sequence, we use DiffuSeq's method (Gong et al., 2022) (see Figure 4) to concatenate source $\mathbf{x_t}$ with target $\mathbf{y_t}$ (see Formula 7), and only partially noise targe sequences and denoise the target with the unnoised source. Word embedding from an existing pre-trained language model is used to convert discrete tokens into embeddings for diffusion continuous attributes at the begining of the forward process, and to turn the continuous representation to tokens at the end of the denoising process.

$$\mathbf{z}_t = \mathbf{x}_t \oplus \mathbf{y}_t, \quad \text{for } t \in [0, T] \tag{7}$$

The training loss function (Formula 8) is based on the variational latent boundary (VLB), aiming to optimize the variational lower bound on the objective function. In this case, the objective function simplifies the KL divergence into end-to-end MSE (Formulas 9 - 11, see Appendix B for more details).

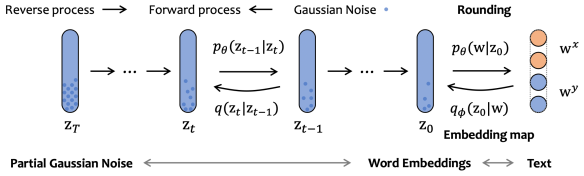


Figure 4: DiffuSeq noising and denoising processes with embedding and rounding through embedding layer, sourced from (Gong et al., 2022)

$$\mathcal{L}_{\text{simple}} \rightarrow \sum_{t=2}^{T} \|\mathbf{y}_0 - \tilde{f}_\theta(\mathbf{z}_t, t)\|^2 + \|\text{EMB}(\mathbf{w}^y) - \tilde{f}_\theta(\mathbf{z}_1, 1)\|^2 + \mathcal{R}(\|\mathbf{z}_0\|^2) \tag{8}$$

5 Experimental

5.1 Datasets

Our experiments are carried out using American sign language (ASL) datasets. The primary dataset used is the ASLG-PC12 corpus (Othman and Jemni, 2012), a substantial parallel corpus that aligns English written texts with ASL glosses. Given that certain experiments involve training models from scratch, the size of the ASLG-PC12 gloss-text parallel corpus is relatively small. To address this issue, data augmentation techniques are employed to generate additional artificial datasets. Artificial glosses are generated from corresponding standard texts from Wikipedia[5], leveraging the features of sign glosses based on sign linguistics. Linguistic features of sign glosses are analyzed from differences between sign language glosses and spoken language, which includes the lack of word inflection, the omission of punctuation and individual words, and syntactic diversity. Consequently, the corresponding heuristics for generating pseudo-glosses from spoken language involve the lemmatization of spoken words, POS-dependent and random word deletion, and random word permutation (Moryossef et al., 2021). This research follows these rules for data augmentation, ensuring the creation of robust pseudo-gloss datasets (See Table 1 for more dataset details).

The overall datasets are divided into train, validation and test datasets with proportion of 70%,

[5] https://www.kaggle.com/datasets/mikeortman/wikipedia-sentences/data

Datasets	Authentic	Artificial
# sentences	87,710	87,710
# words	1,151,110	1,687,804
# glosses	1,029,995	1,079,168
word vocab	22,070	120,273
gloss vocab	16,120	75,069
avg sentence len	13.124	19.243
avg gloss len	11.743	12.304
Train	61,397	61,397
Validation	13,157	13,156
Test	13,157	13,156

Table 1: Dataset used in experiments

15% and 15%. This means that our test dataset is a mixture of both artificial (through data augmentation) and real parallel data. To check how our model performs on real data, we create two test datasets. The first batch is the real dataset in which the parallel data is true glosses with corresponding true standard text translation (**Real** test data); while the second one is all the test data, the mixture of both authentic and artificial test data (**All** test data).

5.2 Evaluation Metrics

To evaluate the performance of different architectures, automatic metrics are used to assess both effectiveness and efficiency. Effectiveness metrics measure how well the predicted outputs align with reference texts, while efficiency metrics focus on the computational cost of model training and inference.

Effectiveness metrics include several measures to assess the quality of generated outputs. **Length** comparison helps determine whether the predictions are appropriately sized relative to the reference texts, highlighting potential tendencies toward overly brief or excessively long outputs. **Accuracy** measures whether the model correctly generates words and tokens, with **token-level accuracy** offering a finer-grained analysis to detect formatting errors. **Levenshtein distance** (Levenshtein et al., 1966) evaluates the syntactical alignment of predicted sequences by counting the minimum number of insertions, deletions, and substitutions required to match reference texts. Additionally, the **BLEU score** (Papineni et al., 2002) assesses the fluency and adequacy of generated sequences by comparing n-gram overlaps with reference texts. We make use of the BLEU score implementation provided by the NLTK package[6].

Efficiency metrics focus on computational performance, specifically **training time** and **inference time**. Training time refers to the total duration required to optimize the model across multiple epochs, excluding validation time. Inference time measures the speed at which a trained model processes new inputs and generates outputs.

5.3 Model Setups

LevT and PointerLevT Both models are encoder-decoder models with 6 layers each, using multi-headed attention, layer normalization, dropout, and embeddings of size 512 from "facebook/mbart-large-cc25". Both models perform up to 10 decoding iterations during inference; they are each trained on 1 NVIDIA A100 GPU.

DiffuSeq DiffuSeq incorporates source texts during training, converting both source and target texts into tensors using a pre-trained BERT tokenizer ("bert-base-uncased") with a vocabulary size of 30522 and $d_{\text{dimension}} = 128$. The noising process applies a square root scheduler $\beta_t = 1 - \sqrt{\frac{t}{T} + 0.0001}$ exclusively to target tensors, using an attention mask (0 for source, 1 for target) to distinguish them. Initially, noise is added uniformly to both tensors, but at the final noising timestep, source tensors are replaced with their original state, resulting in noised target tensors concatenated with unnoised source tensors. This concatenated state serves as the input for the denoising process, which treats both tensors uniformly. At the final denoising step, source tensors are reverted to their initial noising state. The process operates over 2000 timesteps. Multiple seeds are used to select the best outputs through Minimal Bayes Risk (MBR) decoding (Kumar and Byrne, 2004). The model is trained on 1 NVIDIA A100 GPU.

5.4 Baselines

Two types of autoregressive models serve as baseline models, including pre-trained language models, mBART (Chipman et al., 2022) and mT5 (Xue et al., 2020), as well as a small-scale mBART (of the same size as our non-autoregressive models) trained from scratch through knowledge distillation. Besides, a vanilla diffusion model also serves as a baseline for our condition-based diffusion model.

mBART and mT5 are multilingual encoder-decoder models designed for sequence generation

[6]https://www.nltk.org/api/nltk.translate.bleu_score.html

tasks. mBART ("mbart-large-cc25") uses 12 encoder and 12 decoder layers with 250,027 tokens, 1024 embedding size, and GELU activations. mT5 ("mt5-large") has 24 layers each for the encoder and decoder, 250,112 tokens, and similar GELU-based feed-forward layers with relative attention bias. mBART and mT5 are fine-tuned to adapt our dataset, serving as overall baseline models.

KD-mBART applies knowledge distillation using a smaller mBART student model (6 layers, 512 embeddings, 8 attention heads, and 2048 feed-forward dimensions) distilled from a fine-tuned mBART pre-trained teacher model.

The vanilla diffusion employs a 2000-step diffusion process with a square root scheduler for noise generation, defined by $\beta_t = 1 - \sqrt{\frac{t}{T} + 0.0001}$, where $T = 2000$. During denoising, timestep embeddings are integrated into a Transformer encoder and combined with position and input embeddings to incorporate temporal context. Instead of KL divergence, which can lead to instability and complexity, the model adopts MSE for more stable training. To prevent out-of-vocabulary issues, the vocabulary size is set to 13000, and custom word embeddings ($d_{\text{model}} = 128$) are jointly trained with the diffusion loss to optimize computational efficiency and control model size.

The details of models' used in the experiments are displayed in Table 3 in Appendix, including model's number of parameters, batch size of training and test, number of epochs or steps during training, the use of GPU during training and inference.

6 Results

6.1 Effectiveness Discussion

For the two edit-based models, the PointerLevT model demonstrates slightly better performance than the LevT model across most metrics. Specifically, PointerLevT achieves higher BLEU scores compared to LevT. Its word- and token-level accuracies are comparable to those of the fine-tuned mT5 model. Additionally, PointerLevT requires fewer edit operations, as indicated by a lower Levenshtein distance, particularly during inference on the real test set. This suggests that PointerLevT not only generates slightly more accurate sequences but also requires fewer modifications to match reference output. However, the performance difference between the two models remains marginal in general.

For diffusion model, DiffuSeq consistently outperforms vanilla Diffusion and other pre-trained models across multiple evaluation metrics. It achieves significantly higher word accuracy. Token-level accuracy follows a similar trend, highlighting DiffuSeq's superior ability to predict labels and individual tokens more accurately. DiffuSeq also demonstrates a lower Levenshtein Distance, indicating its outputs are closer to the reference texts and require fewer modifications. Additionally, DiffuSeq achieves a BLEU score, far exceeding vanilla Diffusion, The overall performance of effectiveness is comparable to that of the fine-tuned mBART, and also those of two edit-based NAT models.

Although two types of non-autoregressive models result in inferior performance on effectiveness compared to fine-tuned pre-trained models, they outperform their autoregressive counterparts trained from scratch through knowledge distillation. Examples of predicted output generated by each model are displayed in Table 4 in Appendix.

6.2 Efficiency Discussion

For edit-based NAR models, PointerLevT is significantly larger (5.29GB with over 550 million parameters) compared to LevT (1.95GB with 170 million parameters). PointerLevT requires 15.5 hours for training, whereas LevT takes 13.4 hours. However, PointerLevT achieves faster inference, completing the decoding process in 30 seconds, while LevT takes 42 seconds on the real test dataset. Both models were trained with a batch size of 16, but PointerLevT converged in just 10 epochs, whereas LevT required 12 epochs.

Diffusion models have a relatively small model size, with DiffuSeq at 363MB (91,225,274 parameters) and vanilla Diffusion at 334MB (87,336,792 parameters). However, conditioning significantly increases training time. DiffuSeq requires 150 hours to train, compared to just 44 hours for vanilla Diffusion. A similar trend is observed in inference. While DiffuSeq achieves the relatively short inference time per step (24.91s per step for full-text generation), it requires substantially longer—13.84 hours—to reach an optimal balance between accuracy and speed (see Figure 5). This suggests that while conditioning improves performance, it also reduces inference efficiency, which could be a limiting factor in real-time or resource-constrained applications where high effectiveness is needed.

Models	Effectiveness								Efficiency	
	Acc		Acc_token		LevD		BLEU		Train	Inference
	All	Real	All	Real	All	Real	All	Real		
Autoregressive models										
$mBART_{cc25}$	0.68*	0.69*	0.78*	0.82*	7.0*	3.81*	0.48*	0.61*	13.65h	805.25s
$mT5_{large}$	0.54	0.43	0.61	0.50	9.63	8.27	0.28	0.34	8.55h	637.53s
KD-mBART	0.11	0.16	0.38	0.55	14.73	11.36	0.01	0.01	13.75h	23.47s*
Non-autoregressive models										
LevT	0.37	0.37	0.52	0.53	12.08	8.22	0.07	0.13	13.4h	4.2s/42.11s
PointerLevT	0.37	0.41	0.52	0.54	11.11	6.26	0.07	0.15	15.5h	**3.1s/30.97s**
Vanilla Diffusion	0.14	0.17	0.17	0.20	17.51	14.97	0.01	0.01	44h	10.51s/5.84h
DiffuSeq	**0.58**	**0.77**	**0.67**	**0.80**	**11.05**	**3.53**	**0.21**	**0.47**	150h	24.91s/13.84h

Table 2: Models' effectiveness evaluation on all test datasets and real test datasets respectively; efficiency evaluation on real test datasets

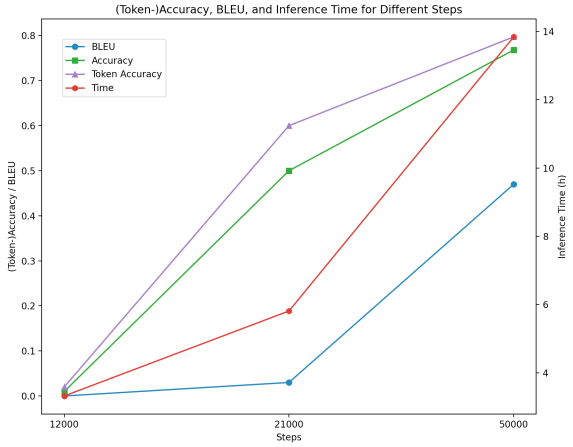


Figure 5: Generated sentences' accuracy, token-accuracy and BLEU in difference steps when DiffuSeq infers

7 Conclusion

This paper investigates two non-autoregressive (NAR) approaches to address the task of translating sign glosses into standard text (gloss2text). While large pretrained models such as mBART and mT5 typically achieve strong performance, our experiments reveal that these NAR models trained with a smaller size of dataset not only match the accuracy of fine-tuned large pre-trained language models but also significantly outperform smaller, distilled models of comparable capacity. In particular, the edit-based NAR strategy strikes a notably favorable balance between accuracy and efficiency, positioning it as a viable alternative to resource-intensive pretrained models. Meanwhile, the conditional diffusion-based approach attains very high accuracy and could benefit from further optimization to enhance its efficiency, making it well-suited for future research.

Moreover, the findings suggest that these methods can be applied to broader text-editing tasks resembling gloss2text, such as grammatical error correction and post-editing. This underscores the potential of novel modeling techniques in both sign language translation and general NLP applications.

Future work may focus on refining both edit-based and diffusion-based NAR models to bolster effectiveness and efficiency across diverse real-world scenarios.

Acknowledgments

We thank the anonymous reviewers for insightful feedback.

References

Nader Akoury, Kalpesh Krishna, and Mohit Iyyer. 2019. Syntactically supervised transformers for faster neural machine translation. *arXiv preprint arXiv:1906.02780*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yu Bao, Hao Zhou, Shujian Huang, Dongqi Wang, Lihua Qian, Xinyu Dai, Jiajun Chen, and Lei Li. 2022. latent-glat: Glancing at latent variables for parallel text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8398–8409.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings

as document identifiers. *Advances in Neural Information Processing Systems*, 35:31668–31683.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*.

Carl Börstell. 2022. Introducing the signglossr package. In *Proceedings of the LREC2022 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources*, pages 16–23.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793.

Hugh A Chipman, Edward I George, Robert E McCulloch, and Thomas S Shively. 2022. mbart: multidimensional monotone bart. *Bayesian Analysis*, 17(2):515–544.

Biao Fu, Peigen Ye, Liang Zhang, Pei Yu, Cong Hu, Yidong Chen, and Xiaodong Shi. 2022. Conslt: A token-level contrastive framework for sign language translation. *arXiv preprint arXiv:2204.04916*.

Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2020. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on computer science and information systems (FedCSIS)*, pages 179–183. IEEE.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. 2022. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. Diffuseq-v2: Bridging discrete and continuous text spaces for accelerated seq2seq diffusion models. *arXiv preprint arXiv:2310.05793*.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. *Advances in neural information processing systems*, 32.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Shankar Kumar and Bill 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.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Jason Lee, Raphael Shu, and Kyunghyun Cho. 2020. Iterative refinement in the continuous space for non-autoregressive neural machine translation. *arXiv preprint arXiv:2009.07177*.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343.

Yifan Li, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Diffusion models for non-autoregressive text generation: A survey. *arXiv preprint arXiv:2303.06574*.

Shan Liu, Yafang Zheng, Lei Lin, Yidong Chen, and Xiaodong Shi. 2023. A novel pos-guided data augmentation method for sign language gloss translation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 392–403. Springer.

Sarfaraz Masood, Adhyan Srivastava, Harish Chandra Thuwal, and Musheer Ahmad. 2018. Real-time sign language gesture (word) recognition from video sequences using cnn and rnn. In *Intelligent Engineering Informatics: Proceedings of the 6th International Conference on FICTA*, pages 623–632. Springer.

Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. 2018. Learning latent permutations with gumbel-sinkhorn networks. *arXiv preprint arXiv:1802.08665*.

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021. Data augmentation for sign language gloss translation. *arXiv preprint arXiv:2105.07476*.

Achraf Othman and Mohamed Jemni. 2012. English-asl gloss parallel corpus 2012: Aslg-pc12. In *5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon LREC*.

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.

Qiu Ran, Yankai Lin, Peng Li, and Jie Zhou. 2021. Guiding non-autoregressive neural machine translation decoding with reordering information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13727–13735.

Razieh Rastgoo, Kourosh Kiani, and Sergio Escalera. 2021. Sign language recognition: A deep survey. *Expert Systems with Applications*, 164:113794.

Richard Sinkhorn and Paul Knopp. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348.

Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit. 2019. Insertion transformer: Flexible sequence generation via insertion operations. In *International Conference on Machine Learning*, pages 5976–5985. PMLR.

MF Tolba and AS Elons. 2013. Recent developments in sign language recognition systems. In *2013 8th International Conference on Computer Engineering & Systems (ICCES)*, pages xxxvi–xlii. IEEE.

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.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in neural information processing systems*, 28.

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*.

Kayo Yin and Jesse Read. 2020. Better sign language translation with stmc-transformer. *arXiv preprint arXiv:2004.00588*.

A Sinkhorn Layer Formulation

The Sinkhorn layer transforms an input matrix $M \in \mathbb{R}^{n \times m}$ into an approximately doubly stochastic matrix P, where each row and column sums to 1. The transformation is defined as:

$$P = \text{Sinkhorn}(M, \tau, T)$$

where:

- M is the input score or cost matrix,
- τ is the temperature parameter controlling sharpness,
- T is the number of Sinkhorn iterations.

Step 1: Initialization (Softmax)

First, apply a softmax-like transformation to ensure non-negativity:

$$K = \exp\left(\frac{M}{\tau}\right)$$

Step 2: Iterative Normalization

For $t = 1, 2, \ldots, T$, perform the following updates:

$$K \leftarrow \frac{K}{\sum_j K_{ij}} \quad \text{(Row normalization)}$$

$$K \leftarrow \frac{K}{\sum_i K_{ij}} \quad \text{(Column normalization)}$$

Step 3: Final Output

After T iterations, the output is:

$$P = K$$

which approximates a doubly stochastic matrix.

B DiffuSeq Objective Equations

There are objective functions for DiffuSeq.

$$\mathbf{z}_t = \mathbf{x}_t \oplus \mathbf{y}_t, \quad \text{for } t \in [0, T] \tag{9}$$

$$\mathcal{L}_{\text{VLB}} = \mathbb{E}_{q(\mathbf{z}_{1:T}|\mathbf{z}_0)} \Bigg[\underbrace{\log \frac{q(\mathbf{z}_T|\mathbf{z}_0)}{p_\theta(\mathbf{z}_T)}}_{\mathcal{L}_T} + \sum_{t=2}^{T} \underbrace{\log \frac{q(\mathbf{z}_{t-1}|\mathbf{z}_0, \mathbf{z}_t)}{p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)}}_{\mathcal{L}_{t-1}} + \underbrace{\log \frac{q_\phi(\mathbf{z}_0|\mathbf{w}^{x \oplus y})}{p_\theta(\mathbf{z}_0|\mathbf{z}_1)}}_{\mathcal{L}_0} - \underbrace{\log p_\theta(\mathbf{w}^{x \oplus y}|\mathbf{z}_0)}_{\mathcal{L}_{\text{round}}} \Bigg] \tag{10}$$

$$\begin{aligned} \min_\theta \mathcal{L}_{\text{simple}} &= \min_\theta \Bigg[\sum_{t=2}^{T} \|\mathbf{z}_0 - f_\theta(\mathbf{z}_t, t)\|^2 \\ &+ \|\text{EMB}(\mathbf{w}^{x \oplus y}) - f_\theta(\mathbf{z}_1, 1)\|^2 - \log p_\theta(\mathbf{w}^{x \oplus y}|\mathbf{z}_0) \Bigg] \\ &\rightarrow \min_\theta [\sum_{t=2}^{T} \|\mathbf{y}_0 - \tilde{f}_\theta(\mathbf{z}_t, t)\|^2 \\ &+ \|\text{EMB}(\mathbf{w}^y) - \tilde{f}_\theta(\mathbf{z}_1, 1)\|^2 \\ &+ \mathcal{R}(\|\mathbf{z}_0\|^2)] \end{aligned} \tag{11}$$

Estimations were conducted using the Machine-Learning Impact calculator presented in (Lacoste et al., 2019).

Models	# Parameters	$Batch_{train}$	$Batch_{test}$	# Epochs/# Steps	GPU_{train}	$GPU_{inference}$
$mBART_{cc25}$	610,851,840	8	16	7	1 L4	1 A100
$mT5_{large}$	1,229,581,312	16	16	9	1 A100	1 A100
KD_{small}	173,207,040	32	16	15	1 L4	1 A100
LevT	172,917,590	16	16	12	1 A100	1 A100
PointerLevT	556,959,062	16	16	10	1 A100	1 A100
Vanilla Diffusion	87,336,792	128	100	600,000	1 A100	1 A100
DiffuSeq	91,225,274	2048	100	50,000	1 A100	1 A100

Table 3: Models' setups

Case Study 1 - Authentic Sentence	
gloss	X-IT BE BEYOND DOUBT THAT PROPOSE LEGISLATION BE ATTACK ON DESC-HUMAN RIGHTS.
reference	it is beyond doubt that the proposed legislation is an attack on human rights.
mBART	it is beyond doubt that the proposed legislation is an attack on human rights.
mT5	it is beyond doubt that the proposed legislation is an attack on human rights.
KD-mBART	it ismena isyonyond theubt thathat theposes legislation beyonttack on humanirrippolicyman rightsights.
LevT	it be beyond doubt that propose legislation be attack on desc-human rights.
PointerLevT	it be beyond doubt that propose legislation be attack on human rights.
DiffuSeq	it is beyond doubt that the proposed legislation is attacks on human rights.
Case Study 2 - Artificial Sentence	
gloss	Thompson Taylor professor political former U.S. science State Oklahoma Representative Carolyn state
reference	Carolyn Thompson Taylor is a former State Representative and professor of political science from the U.S. state of Oklahoma.
mBART	Carolyn Thompson is a former State professor of political science at Oklahoma U.S. Representative State University.
mT5	Carolyn Taylor Thompson is a former professor of political science at the U.S.
KD-mBART	Carol was is of science of.S. statesstiveyn'.
LevT	Thompson Taylor professor political former U.S. science State Oklahoma Representative Carolyn state.
PointerLevT	Carolyn Thompson Taylor former Representative professor U.S. political science State Oklahoma.
Diffusion-LM	of the loss in a and a further - wide development of the industry.
DiffuSeq	thompson is political representative was a science of representative of the professor of u. former. carolyn s the state of state.

Table 4: Predicted Outputs Generated by Models based on one authentic test data and one artificial test data

Exploring the Feasibility of Multilingual Grammatical Error Correction with a Single LLM up to 9B parameters: A Comparative Study of 17 Models

Dawid Wisniewski[1,2], Antoni Solarski[1,3], Artur Nowakowski[1,3],
[1]Laniqo.com, [2]Poznan University of Technology, Poland, [3]Adam Mickiewicz University, Poland
Correspondence: dawid.wisniewski@laniqo.com

Abstract

Recent language models can successfully solve various language-related tasks, and many understand inputs stated in different languages. In this paper, we explore the performance of 17 popular models used to correct grammatical issues in texts stated in English, German, Italian, and Swedish when using a single model to correct texts in all those languages. We analyze the outputs generated by these models, focusing on decreasing the number of grammatical errors while keeping the changes small. The conclusions drawn help us understand what problems occur among those models and which models can be recommended for multilingual grammatical error correction tasks. We list six models that improve grammatical correctness in all four languages and show that Gemma 9B is currently the best performing one for the languages considered[1].

1 Introduction

Grammatical error correction (GEC) is one of the most practical tasks in the Natural Language Processing (NLP) field. Being able to use computers to detect and fix grammatical errors and spelling mistakes is especially beneficial for language learners and professional writers. Recent years have shown great promise in using Large Language Models (LLMs) for various NLP-related tasks, including machine translation, text generation, or text classification. These models, trained on vast amounts of data, learn to predict the most probable continuation of a given sequence. Assuming large corpora are used to train these models, the models should encounter instantiations of various tasks, including questions to be answered followed by their answers, texts expressed in one language followed by their translations, long fragments of texts followed by their summarizations, or grammatically incorrect sentences fixed by some experts as part of the text's continuation. All of these lead to the increasing ability of LLMs to address human requests.

Recently published LLMs are bigger and trained on more data, which enables them to solve more sophisticated tasks with better results. Even though LLMs are currently suboptimal in some tasks, it seems that soon they may become the dominant solution for most NLP-related problems (Kaplan et al., 2020). However, due to LLMs' large sizes and high computational costs, researchers focus on training smaller models of quality similar to bigger ones (Shan, 2024).

In this paper, we aim to explore the abilities of LLMs to solve the GEC task when dealing with several languages at once, including three highly popular ones: English (EN), German (DE), and Italian (IT), as well as a less popular – Swedish (SV). Having one common model for multiple languages may be beneficial in various ways. Using specialized models for each language requires a lot of storage space, which increases the costs of products and hinders the synergy between languages, which is frequently observed in real life. The energy consumption related to using multiple models increases the costs of serving models and impacts our natural environment. To mitigate these issues, smaller models that can be run locally and are more environmentally sustainable may be considered (Schick and Schütze, 2021). These smaller models are easier to control, as they can be loaded, analyzed, and fine-tuned on personal computers, so here, we focus on using single LLMs of moderate size (up to 9B parameters) as they can be loaded on consumer-grade GPUs.

In this paper, we aim to address the following research questions:

RQ1: Which model is the best for multilingual grammar correction considering English, German,

[1]Project supported by grant no. 0311/SBAD/0763 - Mloda Kadra financed by Poznan University of Technology.

Italian, and Swedish?
RQ2: Do models preserve the original input when no errors are present?
RQ3: Which type of prompt is more effective: short and general, or longer and more specific?

2 Related works

The research on the GEC problems has experienced substantial progress in the last years. Recent surveys highlight a shift from traditional rule-based methods, statistical classifiers, and statistical machine translation techniques to neural machine translation, emphasizing the superior performance of neural approaches (Wang et al., 2021; Bryant et al., 2023). However, there are challenges associated with the predominant use of supervised learning in GEC, primarily due to the necessity for annotated datasets. Researchers in the field have reported issues such as annotation inconsistency, human error, and a scarcity of data for languages other than English (Bryant et al., 2023). In this context, leveraging LLM-based GEC with zero or a few examples emerges as a promising solution. Aside from unsupervised approaches, other significant challenges in GEC include multilingualism, low-resource GEC, and the evaluation (Wang et al., 2021; Bryant et al., 2023).

Here, we focus on three main areas relevant to our research: the application of LLMs in GEC, multilingual GEC, and the evaluation of GEC systems.

2.1 Large language models for GEC

In the domain of GEC, LLMs have been utilized in several innovative ways. For instance, LLMs were employed as evaluators of GEC systems (Kobayashi et al., 2024). Large-scale models like GPT-4 (Achiam et al., 2023) have achieved state-of-the-art results in GEC evaluation, demonstrating a higher correlation with human judgments than other methods (Kobayashi et al., 2024). Also, it was proposed to integrate a language model as a grammatical error detection module, thereby enhancing the overall performance of GEC systems (Yasunaga et al., 2021). Nevertheless, LLMs were explored as standalone GEC systems for English (Davis et al., 2024). By employing effective prompting techniques, the authors evaluated seven open-source models and three commercial models. Their findings suggest that LLMs can outperform supervised GEC models on benchmarks annotated with fluency corrections. Furthermore, the study shows that zero-shot generation can be as effective as few-shot approaches. A related study was also conducted for Swedish (Östling et al., 2023), where the properly prompted GPT-3 model heavily outperformed other GEC systems. Similar research, for the same model, was also conducted for English (Loem et al., 2023), where authors focused on the controllability aspect of the prompt-based approach. This versatility highlights the significant potential of LLMs to advance GEC methodologies.

2.2 Multilingual GEC

As multilingualism emerges as a promising direction in GEC, researchers are actively exploring methods to develop and enhance GEC systems for low-resource languages. For instance, certain strategies were proposed for training and fine-tuning GEC models to create a single system capable of handling multiple languages (Rothe et al., 2022; Pająk and Pająk, 2022). Similarly, transfer learning was employed to leverage models trained on high-resource languages (Yamashita et al., 2020). Many LLMs demonstrate at least some degree of multilingual capability, making it logical to explore their potential in the context of multilingual GEC. For instance, the large commercial GPT-3.5 model was examined for its performance in GEC across various languages, yielding promising results. However, the human evaluation revealed that the model encounters difficulties with specific types of errors (Katinskaia and Yangarber, 2024). Collectively, this collection of research demonstrates the feasibility of developing a unified model capable of performing GEC across various languages, including those with limited resources. However, different LLMs, particularly smaller ones, haven't yet been investigated in this context.

2.3 Evaluations

In GEC evaluation, metrics can be broadly categorized into two main types: reference-based and reference-less (Bryant et al., 2023; Wang et al., 2021). While different reference-based metrics are well-studied and widely used, the focus is still on developing improved metrics, particularly those that do not rely on references. It was noted that the existing evaluation methods might stem the progress in the field, as they are tightly coupled with gold-standard references (Rozovskaya and Roth, 2021).

Reference-less metrics for assessing grammat-

ical correctness are very rich. Some of the popular choices use e-rater (Attali and Burstein, 2006) and LanguageTool (Miłkowski, 2010) to calculate grammatical correctness scores. Additionally, an important aspect of GEC systems is the preservation of meaning, for which the US_{IM} (Choshen and Abend, 2018) or BERTScore (Zhang et al., 2020) metrics can be employed. Therefore, further research in GEC evaluation, particularly in the area of semantic faithfulness (meaning preservation), is postulated (Wang et al., 2021). Other aspects are text fluency, for which e.g., the GLEU (Mutton et al., 2007) metric is frequently used, or ensuring that corrections are made using minimal changes rather than reformulations of the whole sentence, which can be analyzed using e.g., Levenshtein distance (Keselj, 2009).

3 Dataset

We use the MultiGED dataset (Volodina et al., 2023) for our experiments. Originally, MultiGED was proposed in 2023 to evaluate the ability of AI systems to identify grammatically incorrect tokens in five languages: English, German, Italian, Czech, and Swedish.

The dataset is a compilation of various datasets that include FCE for English (Yannakoudakis et al., 2011), or Falko-MERLIN for German (Boyd, 2018). As the original dataset represents a binary classification task, where each token in a sentence is classified as correct or not, we preprocessed the dataset using Moses detokenizer (Koehn et al., 2007), to reconstruct entire sentences annotated with the information whether the whole sentence is grammatically correct or not. If any token in the sentence was annotated as incorrect, the full sentence was tagged as incorrect.

There is no golden-standard corrected candidate for a given sentence provided, as the MultiGED dataset is a token classification task. However, following research of (Rozovskaya and Roth, 2021), our goal is to analyze the GEC problem from a broader perspective trying to mitigate annotator bias that may be observed in golden standard corrections.

To achieve this, we use LanguageTool as one of the key tools for scoring the models. As it does not support Czech, we focus on English, German, Italian and Swedish only. To evaluate a wide range of LLMs, we selected the MultiGED's dev set to limit the costs of experiments. This step reduced inference computational power needs, leaving relatively large number of examples for each language considered. The summary of the processed dataset used for further experiments is provided in Table 1.

Language	Total sents	Correct	Tokens per sent
English	2191	906	15.9 (+/- 10.97)
German	2503	619	15.81 (+/- 9.46)
Italian	758	268	11.98 (+/- 7.61)
Swedish	911	199	17.25 (+/- 11.43)
TOTAL	6363	1992	15.59 (+/- 10.21)

Table 1: Dataset summary. **Total sents** column represents the number of sentences for a given language, **Correct** represents the number of sentences marked as grammatically correct, and **Tokens per sent** represents the average number of tokens in a sentence, with standard deviation provided.

4 Models and Methodology

4.1 Models selection

We searched for language models meeting the following criteria: (i) They should fit popular consumer GPUs, thus we set the size limit to 9B parameters, (ii) They, or their base models, should be accompanied by a research paper, (iii) We prefer instruction-following models if present.

As a result, we collected 17 LLMs listed in Table 2. From these, 8 have 7B parameters, 2 have 8B parameters, 3 have 9B parameters, and 4 are smaller than 4B parameters. All models but XGLM and Bloom are instruction-following ones. We decided to consider XGLM and Bloom anyway as their multilingual abilities are strongly underlined in their research papers. Although Karen-strict and Karen-creative are not accompanied by research papers, they are fine-tuned Mistral models that were trained directly to solve the grammatical error correction task.

4.2 Model querying

Every model was queried using three user prompts, which were translated into the target language for inputs other than English (DE, IT, SV):

P1: *Edit the following text for spelling and grammar mistakes:*

P2: *Edit the following text for spelling and grammar mistakes, return only the corrected text:*

P3: *Edit the following text for spelling and grammar mistakes, make minimal changes, and return only the corrected text. If the text is already correct, return it without any explanations:.*

For generation, we use `transformers` library, version 4.42.4. We used each model's `.generate()` function, setting the following parameters:

- `renormalize_logits=False`,
- `do_sample=True`,
- `use_cache=True`,
- `max_new_tokens=256`,
- `repetition_penalty=1.18`,
- `top_k=40`,
- `top_p=0.1`.

The parameter selection is inspired by Karen [2], which is the only model (considered) fine-tuned for GEC. It provides these values as suggestions. We share those values among all models, and leave other generation parameters default.

4.3 Analyzed characteristics

A good GEC system should meet several criteria: (i) it should fix grammatical issues in the text, (ii) it should preserve the meaning of the original text, (iii) it should make possibly small changes to the input text. To gain a better understanding of the LLMs' outputs, we introduced two additional criteria: (iv) it should detect when a given text is correct and should not be changed, and (v) it should preserve the language of the original text. These criteria highlight the inherent dilemma in GEC systems: the trade-off between prioritizing corrections and preserving the original text. The focus on meaning and language preservation stems from the fact that LLMs may generate texts that are not corrected texts (e.g., *There are no errors in this text.*) or may fall back to a different language (e.g., English) if they do not understand a given one. To address these requirements, the following metrics are used, which are calculated for each sentence and then averaged over all examples in a given language:

Req. 1: Grammatical correctness To evaluate language correctness, we use the Python wrapper for LanguageTool [3] (LT), which provides us with information about the list of grammatical errors found in a given text. We use information on the number of errors to evaluate an input sentence s:

$$correctness(s) = \frac{1}{1 + num\_errors(s)}$$

This metric ranges between 0.00 (really bad quality of output) and 1.0 (no grammatical errors found by LT). Good models should increase the value of that metric after correction. The decision to use this metric over other ones is due to the multilingual scenario – LT supports all 4 languages considered, while e.g., e-rater does not.

Req 2: Semantic similarity To assess the semantic similarity between the uncorrected input sentence s_i and the corrected output sentence s_o and detect situations where the text generated is not a correction (e.g., *No errors found*), we calculate three auxiliary metrics, namely: (i) BERTScore (Zhang et al., 2020) similarity calculated using a multilingual BERT [4] (Devlin et al., 2018), (ii) BLEURT (Sellam et al., 2020) similarity calculated using a multilingual model [5], (iii) SentenceBERT (Reimers and Gurevych, 2019) similarity calculated as a cosine similarity between representations of s_i and s_o generated using a multilingual SentenceBERT model [6].

We decided to use three metrics instead of one to make the results more robust and less biased towards one model.

Req 3: Syntactic similarity Apart from reducing the number of errors and preserving meaning, ideally, the model should apply minimal changes to the input text. Similarly to semantic similarity, also here, we use three auxiliary metrics: (i) Levenshtein (edit) score $e()$, which is a transformed edit distance between two sentences s_i and s_o. Levenshtein (edit) distance defines how many changes to the source text s_i should be applied to obtain the target text s_o. The score is calculated as follows:

$$e(s_i, s_o) = \frac{1}{1 + edit\_distance(s_i, s_o)}$$

(ii) GLEU score $g()$, which is one of the most popular metrics for GEC (Mutton et al., 2007). GLEU measures the overlap between the tokens of a hypothesis (here, generated sentence s_o) and a set of references (here, input sentence s_i). The score is calculated using NLTK's (Bird et al., 2009) `sentence_gleu` function with default parameters. (iii) Length difference $d()$, which calculates the difference between s_i and s_o in terms of token numbers. The score is calculated using the following equation, where $cnt()$ returns the number of elements, tokens are generated by $tok()$ function

[2] https://huggingface.co/FPHam/Karen_TheEditor_V2_STRICT_Mistral_7B

[3] https://languagetool.org/

[4] google-bert/bert-base-multilingual-cased

[5] BLEURT-20-D12

[6] sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Name	Huggingface ID
Aya (8B) (Aryabumi et al., 2024)	`CohereForAI/aya-23-8B`
Bloom (7B) (Le Scao et al., 2023)	`bigscience/bloom-7b1`
EuroLLM (1.7B) (Martins et al., 2024)	`utter-project/EuroLLM-1.7B-Instruct`
EuroLLM (9B) (Martins et al., 2024)	`utter-project/EuroLLM-9B-Instruct`
Gemma 2 2B (2B)(Mesnard et al., 2024)	`google/gemma-2-2b-it`
Gemma 2 9B (9B) (Mesnard et al., 2024)	`google/gemma-2-9b-it`
Karen-creative (7B) (Jiang et al., 2023)	`FPHam/Karen_TheEditor_V2_CREATIVE_Mistral_7B`
Karen-strict (7B) (Jiang et al., 2023)	`FPHam/Karen_TheEditor_V2_STRICT_Mistral_7B`
Llama 3.1 (8B) (Touvron et al., 2023)	`meta-llama/Meta-Llama-3.1-8B-Instruct`
Mistral (7B) (Jiang et al., 2023)	`mistralai/Mistral-7B-Instruct-v0.3`
OpenChat 3.5 (7B) (Wang et al., 2023)	`openchat/openchat-3.5-0106`
Phi-3 (3.8B) (Abdin et al., 2024)	`microsoft/Phi-3-mini-4k-instruct`
Qwen 2.5 (7B) (Yang et al., 2024)	`Qwen/Qwen2.5-7B-Instruct`
SmolLM (1.7B) (Allal et al., 2024)	`HuggingFaceTB/SmolLM-1.7B-Instruct`
TowerLLM (7B) (Alves et al., 2024)	`Unbabel/TowerInstruct-7B-v0.2`
XGLM (7.5B) (Lin et al., 2021)	`facebook/xglm-7.5B`
Yi (9B) (Young et al., 2024)	`01-ai/Yi-1.5-9B-Chat`

Table 2: Models selected for the experiment

using `word_tokenize()` method from NLTK, and $max()$ returns the maximum value:

$$d(s_i, s_o) = 1 - \frac{|cnt(tok(s_i)) - cnt(tok(s_o))|}{max(cnt(tok(s_i)), cnt(tok(s_o)))}$$

Although one may expect that tokens may be added or removed (e.g., punctuation marks) after correction, this metric is most sensitive to producing completely different text (e.g., empty output).

Req 4: Keeping correct sentences unchanged Besides correcting errors, LLMs should be able to identify cases, where an input sentence s_i is grammatically correct. The expected behavior in that case is to keep s_i unchanged and return it as the output sentence s_o. Having information about the correctness of sentences from MultiGED, we define true positives, false positives, and false negatives as: the number of examples, where $s_i = s_o$, when s_i is marked correct, the number of examples, where $s_i = s_o$, when s_i is marked incorrect, and the number of examples, where $s_i \neq s_o$ and s_i is marked correct, respectively. Then, we calculate the F_1 score based on these.

Req 5: Language drift Some LLMs have tendencies to produce outputs in a different language than the desired one (Marchisio et al., 2024). This behavior is most frequently observed when a given text expressed in a language other than English is transformed into a text in English without explicitly asking for this kind of switch. For this reason, having an input sentence s_i expressed in language l, that is corrected using a given LLM into a new text s_o, we use the Language Identification tool called LID [7]. LID supports 217 languages and is based on Fasttext (Bojanowski et al., 2016). Knowing what language l a given text s_i is expressed in, we report the probability change:

$$drift(s_o, s_i, l) = P(l|s_o) - P(l|s_i)$$

Since the corrected text should be more probable to be observed, we expect this metric to be $\geq$ 0.0. Values well below 0 mean that the LID tool is more confused about the language after correction, which is not a desired behavior.

5 Results

We calculated values for all the metrics introduced in Section 4.3 using all the prompts described in Section 4.2 for all 17 LLMs considered. Then, we selected the best performing prompt, and used it to verify which LLMs support all the languages, and which LLMs work best for multilingual scenarios as well as for each language separately.

Prompt selection For each metric, each language, and each prompt **P1**, **P2**, **P3**, we calculated the average metric value over all models considered. While the detailed scores can be seen in Table 3, we observe that the best performing prompt is prompt **P3**, which is the longest and most concrete one. This prompt was selected as the best one in 32/36 scenarios considered.

The biggest gain of using the last prompt is seen in the F_1 metric, which checks how well LLMs

[7] https://huggingface.co/facebook/fasttext-language-identification

Prompt [lang]	LT↑	BERT Score↑	SentenceBERT↑	BLEURT↑	Levenshtein↑	Length diff↑	GLEU↑	Language drift↑	Correct (F_1)↑
P1[EN]	0.740	0.750	0.651	0.522	0.092	0.486	0.383	0.049	0.124
P2[EN]	0.804	0.820	**0.739**	0.641	0.169	**0.694**	0.570	**0.035**	0.278
P3[EN]	**0.807**	**0.824**	0.735	**0.647**	**0.197**	0.689	**0.577**	**0.035**	**0.342**
P1[DE]	0.707	0.739	0.659	0.446	0.061	0.461	0.363	**-0.180**	0.103
P2[DE]	0.790	0.794	0.735	0.544	0.090	0.641	0.509	-0.183	0.174
P3[DE]	**0.811**	**0.805**	**0.738**	**0.560**	**0.110**	**0.677**	**0.534**	-0.212	**0.244**
P1[IT]	0.497	0.719	0.624	0.375	0.056	0.443	0.298	**-0.242**	0.065
P2[IT]	0.601	0.775	0.698	0.502	0.111	0.615	0.452	-0.244	0.199
P3[IT]	**0.638**	**0.787**	**0.711**	**0.532**	**0.131**	**0.658**	**0.487**	-0.262	**0.249**
P1[SV]	0.420	0.728	0.646	0.395	0.048	0.444	0.306	-0.284	0.081
P2[SV]	0.524	0.792	0.716	0.512	0.094	0.655	0.486	-0.286	0.187
P3[SV]	**0.552**	**0.804**	**0.725**	**0.539**	**0.113**	**0.682**	**0.518**	**-0.283**	**0.248**

Table 3: Prompt selection vs. metrics for each language. For each language and prompt, a given row represents scores averaged over all 17 models considered. Prompt identifiers are the same as introduced in Section 4.2.

Model	LT↑	BERT Score↑	SentenceBERT↑	BLEURT↑	Levenshtein↑	Length diff↑	GLEU↑	Language drift↑	Correct (F_1)↑
Aya	0.900	0.920	0.935	0.772	0.259	**0.928 (3)**	**0.797 (3)**	0.015	**0.481 (3)**
BLOOM	0.184	0.516	0.028	0.249	0.001	0.056	0.032	-0.693	0.000
EuroLLM (1.7B)	0.912	0.849	0.848	0.639	0.119	0.854	0.583	0.011	0.223
EuroLLM (9B)	0.915	0.888	0.902	0.712	0.218	0.790	0.702	**0.029 (2)**	0.437
Gemma (2B)	**0.940 (2)**	**0.920 (3)**	**0.941 (3)**	0.755	0.121	0.926	0.774	0.012	0.442
Gemma (9B)	**0.948 (1)**	**0.937 (1)**	**0.954 (1)**	**0.774 (3)**	0.136	**0.942 (1)**	**0.814 (2)**	0.017	**0.560 (1)**
Karen (creative)	0.705	0.895	0.933	0.643	**0.276 (2)**	0.924	0.687	-0.297	0.457
Karen (strict)	0.662	0.879	0.913	0.575	**0.268 (3)**	0.906	0.622	-0.432	0.463
Llama 3.1	**0.937 (3)**	0.887	0.894	0.709	0.148	0.853	0.700	**0.032 (1)**	0.255
Mistral	0.710	0.810	0.847	0.532	0.044	0.655	0.489	-0.151	0.022
OpenChat	0.934	0.919	0.932	**0.775 (2)**	0.242	0.925	0.793	**0.018 (3)**	0.420
Phi	0.428	0.655	0.533	0.285	0.006	0.230	0.142	-0.065	0.002
Qwen 2.5	0.896	**0.935 (2)**	**0.952 (2)**	**0.805 (1)**	**0.298 (1)**	**0.939 (2)**	**0.845 (1)**	0.008	**0.545 (2)**
SmolLM	0.117	0.536	0.060	0.253	0.001	0.065	0.031	-0.695	0.000
TowerLLM	0.653	0.832	0.840	0.503	0.141	0.804	0.500	-0.384	0.251
XGLM	0.366	0.513	0.095	0.168	0.007	0.167	0.044	-0.508	0.000
Yi	0.729	0.788	0.754	0.532	0.058	0.536	0.436	0.015	0.047

Table 4: Scores macro-averaged over languages (EN, DE, IT, SV).

Model	Aggregated rank	Rank EN	Rank DE	Rank IT	Rank SV	Supports all langs	Improves LT on
Gemma (9B)	1	4	6	1	1	YES	EN, DE, IT, SV
Qwen 2.5	2	2	1	3	2	YES	EN, DE, IT
Aya	3	7	2	2	5	YES	EN, DE, IT
Gemma (2B)	4	7	4	4	3	YES	EN, DE, IT, SV
OpenChat	5	5	2	4	4	YES	EN, DE, IT, SV
EuroLLM (9B)	6	6	8	7	7	YES	EN, DE, IT, SV
Karen (creative)	6	2	5	11	5	NO	EN, DE
Llama 3.1	8	9	7	9	8	YES	EN, DE, IT, SV
Karen (strict)	9	1	8	7	10	NO	EN, DE
EuroLLM (1.7B)	10	12	11	6	9	YES	EN, DE, IT, SV
TowerLLM	11	11	10	10	14	NO	EN, DE
Mistral	12	10	12	13	12	NO	EN, DE
Yi	13	13	13	12	11	YES	EN, DE
Phi	14	14	14	14	13	NO	–
BLOOM	15	15	16	16	15	NO	–
XGLM	15	16	15	15	16	NO	–
SmolLM	17	16	16	17	17	NO	–

Table 5: Aggregated ranks with ties represented as the same positions. Models in bold support all languages and improve correctness (LT) metric on each language. Since Aya, Qwen 2.5, and Karen are very good on some languages (Karen is top-scored on English, Aya is second on Italian and German, and Qwen 2.5 is top-scored for German), they outperform some other models marked in bold, which support all languages but with worse quality.

keep correct sentences unchanged. This observation agrees with the intuition – in prompt **P3**, we explicitly tell those models not to change the input in the last scenario. The significant increase of F_1 score for each language shows that the model understands this kind of query. In the subsequent analyses, prompt **P3** is used.

Language support analysis Several LLMs struggle with handling languages other than English, often producing outputs (or considerable fragments of outputs) in English. Models like Bloom and SmolLM exhibit the highest drifts, converting a significant portion of German, Italian, and Swedish texts into English. Other models, such as XGLM, Karen-creative, Karen-strict, TowerLLM, Mistral, and Phi-3, also show considerable drift, with each producing outputs in a different language in at least a quarter of the cases. Consequently, these models are not considered effective for multilingual grammatical error correction. Detailed per-language results can be found in Appendix, in Tables 9, 10, 11, 12. The aggregated scores can be found in Table 4.

Out of 17 LLMs considered, 9 of them (Aya, EuroLLM 1.7B, EuroLLM 9B, Gemma 2B and 9B, LLama 3.1, Openchat 3.5, Qwen 2.5, and Yi) produce outputs in all languages considered in a vast majority of cases. Mistral and Phi-3 fail to produce texts in one language - Italian, while the remaining models have problems with at least two languages. Since not all LLMs support all languages, we repeated the prompt quality analysis considering only Aya, EuroLLMs, Gemmas, LLama 3.1, Openchat 3.5, Qwen 2.5, and Yi. As a result, we confirmed the previous conclusion – the **P3** prompt is the best one in this scenario in 32/36 cases. The details on this analysis can be found in Appendix, Table 8.

Ranking models As our goal is to identify models that correct texts with the highest quality (maximizing correctness metric), preserving the meaning of the original text (maximizing metrics based on BERT Score, SentenceBERT and BLEURT), applying possibly small changes (maximizing metrics based on Levenshtein, Length difference, and GLEU), and estimating the ability not to change a given text when it is correct (maximizing F_1 metric), we analyzed those metrics per language and then aggregated them to obtain a general overview of these models. We left language drift aside, as it was primarily used to filter out LLMs not supporting all languages considered.

To reach this goal, we followed a three-step analysis consisting of metric calculation, per-metric ranking creation, and global ranking creation:

(i) Metric calculation – First, we calculated metric values for each model and each language considered. These are provided in Appendix A, Table 9 for English, and Tables 10, 11, 12 for German, Italian, and Swedish, respectively. These metrics were then macro-averaged to obtain a general, language-agnostic view of these models assigning each language the same weight. The aggregated scores are presented in Table 4.

(ii) Per-metric ranking creation – For each metric considered, we ranked each model according to the metric value. We applied this procedure both to per-language scores described in Tables 9- 12, and for averaged metrics from Table 4. For brevity, in this paper, we explicitly present only the rankings for the aggregated metrics in Table 6 as they can be easily created by sorting each language model according to a given metric.

(iii) Global ranking creation – Finally, in order to get a single rank assigned to each model, we perform rank aggregation based on rankings introduced in the previous paragraph using the Borda method (McLean, 1990). In the first iteration, Borda aggregation is calculated separately for semantic metrics (BERTScore, Sentence BERT, BLEURT), as well as for syntactic ones (GLEU, length diff, Levensthein). Then, we applied the Borda aggregation on LanguageTool ranks, Correct F_1 ranks and newly calculated semantic and syntactic ranks. This two-step scenario is required to give each perspective (text accuracy, semantics, syntax, text preservation) the same weight, when having multiple metrics for some of them (here, semantics and syntax). The results of this step are present in Table 5.

Rankings overview The rankings presented in Table 5 show that Gemma 9B, Qwen 2.5, and Aya are the top-ranked models when considering all languages at once. A per-language analysis shows that for English Karen models (explicitly fine-tuned for the GEC task) take the lead, with Qwen 2.5 ranked an par with Karen (strict) as the second one. Qwen and Aya are good choices for German, followed by OpenChat and Gemmas, while for Italian Gemma 9B is the best, and Aya and Qwen are the runner ups. For Swedish both Gemmas, Qwen 2.5 and Openchat work really well. Overall, taking into considerations rankings created and requirements

Model	LT↑	BERT Score↑	SentenceBERT↑	BLEURT↑	Levenshtein↑	Length diff↑	GLEU↑	Language drift↑	Correct (F_1) ↑
Aya	7	4	4	4	4	**3**	**3**	6	**3**
BLOOM	16	16	17	16	16	17	16	16	15
EuroLLM (1.7B)	6	10	10	9	11	8	10	8	11
EuroLLM (9B)	5	7	8	6	6	11	6	**2**	7
Gemma (2B)	**2**	**3**	**3**	5	10	4	5	7	6
Gemma (9B)	**1**	**1**	**1**	**3**	9	**1**	**2**	4	**1**
Karen (creative)	11	6	5	8	**2**	6	8	12	5
Karen (strict)	12	9	7	10	**3**	7	9	14	4
Llama 3.1	**3**	8	9	7	7	9	7	**1**	9
Mistral	10	12	11	11	13	12	12	11	13
OpenChat	4	5	6	**2**	5	5	4	**3**	8
Phi	14	14	14	14	15	14	14	10	14
Qwen 2.5	8	**2**	**2**	**1**	**1**	**2**	**1**	9	**2**
SmolLM	17	15	16	15	17	16	17	17	16
TowerLLM	13	11	12	13	8	10	11	13	10
XGLM	15	17	15	17	14	15	15	15	17
Yi	9	13	13	12	12	13	13	5	12

Table 6: Ranks generated based on Table 4

of text quality improvement and all language support, Gemma 9B and 2B, EuroLLMs 9B and 2B as well as OpenChat and Llama 3.1 are good options. While EuroLLMs are ranked behind Aya and Qwen in some cases, they support all languages and improve grammatical correctness in each language.

6 Discussion

The rankings presented in Table 5 show that Gemma 9B is currently the best model overall for the GEC task. Even if it is not fine-tuned, it achieves outstanding results. This model can be a good foundation model for a supervised fine-tuning process. Similarly, Aya is also very good, but its limited quality on Swedish makes Gemma 9B a better choice. Qwen 2.5 is an interesting case, as it is one of the best performing models (ranked second overall, the best for German, second on English and Swedish, and third on Italian). However, it does not decrease the number of errors on Swedish introducing a slight decrease on the language tool metric. At the same time, it is scored very high on semantic and syntactic similarity metrics as well as F_1 score, which means that this model tends to copy Swedish texts rather than correct them. The high score of Karen models, outperforming Mistral (their base model) proves that such fine-tuning may further increase the quality of models.

Model size vs. quality Four models considered are much smaller than the rest – Phi-3 (3.7B), Gemma 2B (2B), SmolLM (1.7B), and EuroLLM (1.7B). While Phi-3's performance is mediocre, SmolLM behaves badly on the GEC task. One may think that it is due to an insufficient number of parameters. However, Gemma 2B (ranked foruth overall) and EuroLLM 1.7B (ranked 10th), which are of similar size to SmolLM (ranked 17) present outstanding performance considering their sizes. Even though Gemma 2B and EuroLLM 1.7B are very small, they understand all four languages. This proves that appropriate training (Gemma 2B is a distilled version of larger Gemma) and appropriate data (EuroLLMs are European language-focused) are more crucial than the model size itself.

Recurring problems Several recurring issues have been identified in the results produced by LLMs. The most popular examples are listed in Table 7. Additionally, we listed contrastive analysis of examples rated high on one metric by one model and low by another one. These examples are provided in Appendix A, Tables 13-15, representing one semantic metric, one syntactic metric, and LanguageTool. Firstly, LLMs may produce empty results (e.g., for 36% of German examples processed with XGLM). Secondly, LLMs may generate new text instead of correcting the existing input (e.g., the majority of XGLM, SmolLM, Bloom outputs). This behavior is also occasionally observed for other models in case of short inputs provided (e.g., TowerLLM). It is a consequence of the inherent design of LLMs, which are trained on extensive corpora to predict the next token. Thirdly, LLMs sometimes provide explanations or justifications for corrections rather than simply presenting the corrected text (most frequently observed for Yi and Mistral). These models tend to be verbose and informative, even when explicitly instructed otherwise. An important issue observed (e.g., for Yi) is language mixing, where an actual correction in a given language is accompanied by an English comment (even if the prompt is formulated in language other than English). This behavior may be one of the main reasons for a decrease of the language drift metric values and limited languages support observed. Other kinds of recurring problems are: making paraphrases instead of small corrections,

copying prompts (e.g., TowerLLM), answering in a different language (e.g., TowerLLM), and ignoring the input provided (e.g., Phi-3, frequently generating: *you haven't provided a specific passage*).

Result stability Due to the non-deterministic nature of LLM outputs, we also assessed the stability of the generation process to ensure that our results were consistent and that the rankings obtained in repeated runs of the experiments remained stable. For that, we selected the best-performing model (Gemma 9B) and a randomly chosen one (TowerLLM) and reran the experiments five times, maintaining all original parameters. Although minor differences were observed, their magnitudes were negligible. Across both models and all prompts, the variation in Language Tool scores did not significantly exceed 0.001. The largest difference for the Gemma 9B model was 0.00037, obtained on the Swedish dataset with prompt **P1**. While for the TowerLLM model, it was 0.00129, obtained on the Italian dataset with prompt **P3**. Similar magnitudes of deviations were observed for other metrics. Based on these findings, we conclude that the results are both reliable and stable across runs.

7 Karen analysis: corrections vs preservation

In the field of GEC, there is an inherent dilemma: should a model prioritize making corrections or preserving the original text with minimal changes? When aiming to preserve the original text, it's crucial to consider both semantic faithfulness and syntax preservation. This challenge is particularly pronounced for LLMs, which tend to over-correct, or even paraphrase provided texts. The authors of Karen models addressed this by offering two versions: "strict" and "creative." When focusing exclusively on the English language, the Karen-strict model excels by making minimal changes and maintaining the highest similarity to the original text. It is ranked above the creative version across all metrics related to both semantic and syntactic preservation. Despite being ranked 8th in the LT score for English, the model still achieved a high score of 0.938 (with the best model scoring 0.963). Conversely, the Karen-creative model produces higher quality output (with an LT score of 0.946), but the differences between the original and corrected texts are more significant. It still maintains high semantic and syntactic preservation, especially when compared to other LLMs. This performance can be attributed to the Karen models being fine-tuned specifically for the GEC task.

8 Conclusions

In this paper, we analyzed whether popular LLMs not larger than 9B parameters are able to handle grammatical error correction in a multilingual manner and proposed a framework for referenceless GEC LLM comparison. We found that Gemma (2B and 9B), EuroLLM (1.7B and 9B), Openchat 3.5, and LLaMA 3.1 are able to handle all analyzed languages (EN, DE, IT, SV) with good quality. Since Gemma 9B is the top-ranked model in the multilingual scenario and the language-averaged LT score and F_1-correct input preservation are the highest for it, it is the recommended model to use (answer to **RQ1**). Also, its smaller sibling – Gemma 2B – performs very well. We found out that some models (Gemmas, EuroLLM 9B, Qwen, Karens, Aya, and OpenChat) preserve the original text well if no error is introduced. Comparing fine-tuned models (Karen strict and Karen creative) with their base Mistral model, we see that the fine-tuned models excel in terms of overall correction quality, which shows the importance of fine-tuning to prevent hallucinations in LLMs. Thus, **RQ2** can be answered positively. Finally, we found that the longest, most concrete prompt is the best performing one overall, which answers **RQ3**.

We made the corrections generated using all prompts and models publicly available online [8].

9 Limitations

Our rankings focus on relatively small LLMs. However, there are much larger models trained on more data. We suppose that these may exhibit better performance, however, they are more costly, and harder to control, as most frequently they are hidden behind remote APIs. Thus, the conclusions drawn here refer to models of moderate size (up to 9B parameters), and bigger models may perform better. Additionally, selecting LanguageTool as the source of error information may introduce some bias towards errors detected by LanguageTool. It may be that some kinds of problems are overlooked, and some false positives are produced despite the maturity of this tool. However, LT is widely used for GEC evaluation.

[8] https://github.com/laniqo-public/grammar-data-mtsummit25

Language	Model	Type	Text generated	Times observed in outputs
German	XGLM	End of sequence token	<lim_end|>	902
Italian	XGLM	End of sequence token	<lim_end|>	287
English	XGLM	End of sequence token	<lim_end|>	276
English	Yi	Comments added	"grammatically correct" in output text	327
English	Mistral	Comments added	"grammatically correct" in output text	108
English	Qwen	Comments added	"grammatically correct" in output text	4
English	Yi	Comments added	"the text is" in output text	65
English	Tower	Comments added	"the text is" in output text	4
English	mistral	Comments added	"no corrections needed" in output text	279
English	Yi	Comments added	"no corrections needed" in output text	65
English	XGLM	Unrelated text generated	"Delete all" as the beginning of the generated text	1420
English	SmolLM	Unrelated text generated	"the first step in writing a research paper" as the beginning of the generated text	682
German	Bloom	Unrelated text generated + language drift	"introduction the use of the internet has become a commonplace part" as the beginning of the generated text	204
German	SmolLM	Unrelated text generated + language drift	"the 1960s were an era of great change in the" as output starter	556
English	Phi-3	no input detected	"you haven't provided a specific passage" in output text	61
German	TowerLLM	prompt copying	"korrigieren sie den folgenden text auf rechtschreib- und grammatikfehler, nehmen sie nur minimale änderungen vor und senden sie nur den korrigierten text zurück. wenn der text bereits korrekt ist, senden sie ihn ohne erklärungen zurück" in the generated text	36
German	TowerLLM	prompt copying + translating	"please correct the following text for spelling and grammar errors, make only minimal changes if necessary, and send back just the corrected text" in generated text	27
German	Yi	language mixing	"Here is the corrected" text generated before the actual correction	30
Swedish	Gemma 2B	no input detected + Language drift	please provide the text you would like me to correct!	9
Italian	Gemma 9B	no input detected	per favore, fornisci il testo che vuoi correggere.	6
German	Gemma 9B	no input detected	bitte geben sie den text ein, den ich korrigieren soll.	7
German	Gemma 9B	no input detected	bitte geben sie mir den text zum korrekturlesen.	3
Swedish	Gemma 9B	no input detected	vänligen ge mig texten du vill att jag ska redigera!	3

Table 7: Examples of recurrent problems for given models and languages.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam et al. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Leandro von Werra, and Thomas Wolf. 2024. Smollm - blazingly fast and remarkably powerful.

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.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v.2. *The Journal of Technology, Learning and Assessment*, 4(3).

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Adriane Boyd. 2018. Using wikipedia edits in low resource grammatical error correction. In *Proceedings of the 4th Workshop on Noisy User-generated Text, NUT@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 79–84. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical Error Correction: A Survey of the State of the Art. *Computational Linguistics*, 49(3):643–701.

Leshem Choshen and Omri Abend. 2018. Reference-less measure of faithfulness for grammatical error correction. 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 124–129, New Orleans, Louisiana. Association for Computational Linguistics.

Christopher Davis, Andrew Caines, O Andersen, Shiva Taslimipoor, Helen Yannakoudakis, Zheng Yuan, Christopher Bryant, Marek Rei, and Paula Buttery. 2024. Prompting open-source and commercial language models for grammatical error correction of English learner text. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11952–11967, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Anisia Katinskaia and Roman Yangarber. 2024. GPT-3.5 for grammatical error correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7831–7843, Torino, Italia. ELRA and ICCL.

Vlado Keselj. 2009. Speech and language processing (second edition) daniel jurafsky and james h. martin (stanford university and university of colorado at boulder) pearson prentice hall, 2009, hardbound, ISBN 978-0-13-187321-6. *Comput. Linguistics*, 35(3):463–466.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. Large language models are state-of-the-art evaluator for grammatical error correction. *Preprint*, arXiv:2403.17540.

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.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

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*.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of gpt-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. *Preprint*, arXiv:2305.18156.

Kelly Marchisio, Wei-Yin Ko, Alexandre Bérard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in llms. *CoRR*, abs/2406.20052.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno Miguel Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, M. 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. *CoRR*, abs/2409.16235.

Iain McLean. 1990. The borda and condorcet principles: three medieval applications. *Social Choice and Welfare*, 7(2):99–108.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Marcin Miłkowski. 2010. Developing an open-source, rule-based proofreading tool. *Software Practice and Experience*, 40:543–566.

Andrew Mutton, Mark Dras, Stephen Wan, and Robert Dale. 2007. GLEU: automatic evaluation of sentence-level fluency. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Krzysztof Pająk and Dominik Pająk. 2022. Multilingual fine-tuning for grammatical error correction. *Expert Systems with Applications*, 200:116948.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2022. A simple recipe for multilingual grammatical error correction. *Preprint*, arXiv:2106.03830.

Alla Rozovskaya and Dan Roth. 2021. How good (really) are grammatical error correction systems? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2686–2698, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. *Preprint*, arXiv:2009.07118.

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.

Richard Shan. 2024. Language artificial intelligence at a crossroads: Deciphering the future of small and large language models. *Computer*, 57(8):26–35.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Elena Volodina, Christopher Bryant, Andrew Caines, Orphée De Clercq, Jennifer-Carmen Frey, Elizaveta Ershova, Alexandr Rosen, and Olga Vinogradova. 2023. Multiged-2023 shared task at nlp4call: Multilingual grammatical error detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 1–16.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. A comprehensive survey of grammatical error correction. *ACM Trans. Intell. Syst. Technol.*, 12(5).

Ikumi Yamashita, Satoru Katsumata, Masahiro Kaneko, Aizhan Imankulova, and Mamoru Komachi. 2020. Cross-lingual transfer learning for grammatical error correction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4704–4715, Barcelona, Spain (Online). International Committee on Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*,

pages 180–189. The Association for Computer Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2021. Lm-critic: Language models for unsupervised grammatical error correction. *Preprint*, arXiv:2109.06822.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Robert Östling, Katarina Gillholm, Murathan Kurfalı, Marie Mattson, and Mats Wirén. 2023. Evaluation of really good grammatical error correction. *Preprint*, arXiv:2308.08982.

Sustainability statement The experiments were run on a single PC machine equipped with one GeForce 4090 RTX GPU. The correction lasted about 24 hours, with additional one hour for metric evaluation. Based on these, the CO2 impact calculated using `https://mlco2.github.io/impact/` tool is equal to 3.24 kg of CO2eq.

A Detailed analysis and scores

Prompt [lang]	LT↑	BERT Score↑	SentenceBERT↑	BLEURT↑	Levenshtein↑	Length diff↑	GLEU↑	Language drift↑	Correct (F_1)↑
P1[EN]	0.821	0.779	0.757	0.545	0.082	0.514	0.406	**0.067**	0.098
P2[EN]	**0.937**	0.891	0.902	0.732	0.184	**0.858**	0.705	0.037	0.324
P3[EN]	0.923	**0.899**	**0.904**	**0.746**	**0.236**	**0.858**	**0.724**	0.033	**0.448**
P1[DE]	0.755	0.766	0.751	0.481	0.044	0.456	0.380	-0.007	0.066
P2[DE]	0.906	0.862	0.881	0.653	0.094	0.775	0.642	-0.010	0.190
P3[DE]	**0.928**	**0.888**	**0.905**	**0.701**	**0.132**	**0.848**	**0.709**	**0.001**	**0.319**
P1[IT]	0.672	0.760	0.716	0.447	0.065	0.449	0.363	**0.042**	0.082
P2[IT]	0.833	0.855	0.843	0.650	0.146	0.746	0.616	0.010	0.283
P3[IT]	**0.906**	**0.884**	**0.887**	**0.710**	**0.183**	**0.830**	**0.686**	0.027	**0.372**
P1[SV]	0.630	0.769	0.755	0.467	0.045	0.450	0.366	**0.029**	0.068
P2[SV]	0.795	0.878	0.882	0.668	0.123	0.813	0.676	0.005	0.262
P3[SV]	**0.849**	**0.904**	**0.909**	**0.719**	**0.160**	**0.883**	**0.744**	0.009	**0.377**

Table 8: Prompt selection vs. metrics for each language, considering only models that support all languages. For each language and prompt, a given row represents scores averaged over all 7 models supporting all four languages considered. Prompt identifiers are the same as introduced in Section 4.2.

Model	LT (0.754)	BERT Score	SentenceBERT	BLEURT	Levenshtein	Length diff	GLEU	Language drift	Correct (F_1)
Aya	0.919	0.944	0.957	0.826	0.365	0.954	0.857	0.019	0.568
BLOOM	0.558	0.550	0.027	0.320	0.001	0.065	0.033	**0.085 (1)**	0.000
EuroLLM (1.7B)	0.936	0.810	0.784	0.594	0.061	0.802	0.442	0.051	0.075
EuroLLM (9B)	0.939	0.929	0.942	0.808	0.356	0.929	0.822	0.032	0.571
Gemma (2B)	**0.963 (1)**	0.924	0.945	0.766	0.137	0.935	0.771	0.027	0.522
Gemma (9B)	**0.949 (3)**	0.946	0.961	0.791	0.161	0.950	0.831	0.026	0.638
Karen (creative)	0.946	**0.946 (3)**	**0.965 (3)**	**0.838 (3)**	**0.424 (3)**	**0.958 (3)**	**0.871 (3)**	0.029	**0.646 (3)**
Karen (strict)	0.938	**0.956 (2)**	**0.968 (2)**	**0.857 (2)**	**0.471 (1)**	**0.959 (2)**	**0.896 (2)**	0.015	**0.680 (1)**
Llama 3.1	0.939	0.919	0.928	0.778	0.250	0.913	0.792	0.032	0.429
Mistral	**0.958 (2)**	0.838	0.854	0.653	0.057	0.712	0.554	0.069	0.043
OpenChat	0.948	0.932	0.951	0.803	0.309	0.948	0.821	0.034	0.515
Phi	0.612	0.662	0.589	0.373	0.009	0.250	0.161	0.071	0.007
Qwen 2.5	0.935	**0.961 (1)**	**0.970 (1)**	**0.861 (1)**	**0.452 (2)**	**0.960 (1)**	**0.913 (1)**	0.004	**0.680 (2)**
SmolLM	0.233	0.543	0.045	0.283	0.001	0.067	0.032	**0.077 (2)**	0.000
TowerLLM	0.894	0.882	0.853	0.723	0.248	0.834	0.687	0.049	0.402
XGLM	0.281	0.535	0.049	0.229	0.005	0.149	0.051	-0.096	0.000
Yi	0.774	0.726	0.699	0.490	0.036	0.330	0.271	**0.076 (3)**	0.032

Table 9: Scores for English. Models marked in bold improve LanguageTool score in comparison to the text before correction (0.754). Numbers (1), (2), (3) represent the top-3 ranked models on a given metric.

Model	LT (0.74)	BERT Score	SentenceBERT	BLEURT	Levenshtein	Length diff	GLEU	Language drift	Correct (F_1)
Aya	**0.971 (1)**	0.907	0.938	**0.752 (3)**	0.169	0.921	0.769	0.006	0.376
BLOOM	0.159	0.508	0.040	0.218	0.001	0.055	0.036	-0.978	0.000
EuroLLM (1.7B)	0.923	0.863	0.883	0.671	0.099	0.889	0.636	0.002	0.180
EuroLLM (9B)	0.934	0.888	0.923	0.706	0.190	0.803	0.722	**0.009 (2)**	**0.514 (1)**
Gemma (2B)	0.951	**0.916 (3)**	**0.943 (3)**	0.739	0.105	0.923	0.776	0.006	0.435
Gemma (9B)	0.940	**0.922 (2)**	**0.949 (1)**	0.732	0.097	**0.930 (2)**	**0.787 (2)**	0.006	0.351
Karen (creative)	0.936	0.913	0.936	0.708	**0.274 (1)**	**0.930 (1)**	0.758	-0.164	**0.491 (2)**
Karen (strict)	0.945	0.863	0.882	0.534	**0.222 (2)**	0.869	0.580	-0.471	**0.455 (3)**
Llama 3.1	**0.958 (2)**	0.898	0.921	0.730	0.121	0.890	0.739	**0.014 (1)**	0.204
Mistral	0.895	0.833	0.888	0.597	0.042	0.695	0.580	-0.020	0.013
OpenChat	**0.954 (3)**	0.915	0.937	**0.762 (2)**	0.179	0.924	**0.786 (3)**	**0.008 (3)**	0.354
Phi	0.627	0.663	0.582	0.313	0.005	0.226	0.147	-0.061	0.000
Qwen 2.5	0.940	**0.923 (1)**	**0.948 (2)**	**0.766 (1)**	**0.205 (3)**	**0.927 (3)**	**0.816 (1)**	0.004	0.449
SmolLM	0.221	0.539	0.088	0.242	0.001	0.066	0.033	-0.978	0.000
TowerLLM	0.949	0.845	0.894	0.469	0.126	0.888	0.508	-0.524	0.321
XGLM	0.699	0.523	0.083	0.130	0.006	0.155	0.043	-0.407	0.000
Yi	0.780	0.757	0.708	0.448	0.023	0.421	0.353	-0.051	0.006

Table 10: Scores for German. Models marked in bold improve LanguageTool score in comparison to the text before correction (0.74). Numbers (1), (2), (3) represent the top-3 ranked models on a given metric.

Model	LT (0.851)	BERT Score	SentenceBERT	BLEURT	Levenshtein	Length diff	GLEU	Language drift	Correct (F_1)
Aya	**0.967 (2)**	**0.917 (3)**	**0.937 (3)**	**0.784 (3)**	**0.287 (2)**	0.921	**0.784 (3)**	0.022	**0.541 (2)**
BLOOM	0.008	0.495	0.028	0.230	0.001	0.040	0.025	-0.934	0.000
EuroLLM (1.7B)	0.900	0.855	0.847	0.640	0.173	0.844	0.604	0.011	0.314
EuroLLM (9B)	0.889	0.852	0.851	0.637	0.156	0.663	0.590	**0.048 (1)**	0.305
Gemma (2B)	**0.958 (3)**	0.914	0.936	0.760	0.125	**0.923 (3)**	0.753	**0.014 (3)**	0.384
Gemma (9B)	**0.970 (1)**	**0.938 (1)**	**0.954 (1)**	**0.803 (2)**	0.164	**0.944 (1)**	**0.814 (2)**	0.029	**0.644 (1)**
Karen (creative)	0.386	0.829	0.903	0.436	0.147	0.870	0.467	-0.645	0.245
Karen (strict)	0.420	0.847	0.900	0.480	0.205	0.884	0.521	-0.607	0.335
Llama 3.1	0.924	0.844	0.833	0.629	0.115	0.750	0.577	**0.046 (2)**	0.167
Mistral	0.530	0.775	0.816	0.429	0.046	0.612	0.391	-0.353	0.022
OpenChat	0.950	0.908	0.925	0.776	**0.262 (3)**	0.913	0.766	0.023	0.457
Phi	0.315	0.640	0.437	0.197	0.005	0.215	0.124	-0.253	0.000
Qwen 2.5	0.912	**0.928 (2)**	**0.946 (2)**	**0.814 (1)**	**0.291 (1)**	**0.929 (2)**	**0.819 (1)**	0.020	**0.518 (3)**
SmolLM	0.006	0.525	0.050	0.247	0.001	0.051	0.024	-0.934	0.000
TowerLLM	0.621	0.833	0.830	0.536	0.169	0.806	0.515	-0.315	0.283
XGLM	0.413	0.478	0.139	0.097	0.015	0.230	0.040	-0.653	0.000
Yi	0.682	0.801	0.751	0.549	0.072	0.584	0.469	0.027	0.022

Table 11: Scores for Italian. Models marked in bold improve LanguageTool score in comparison to the text before correction (0.851). Numbers (1), (2), (3) represent the top-3 ranked models on a given metric.

Model	LT (0.802)	BERT Score	SentenceBert	BLEURT	Levenshtein	Length diff	GLEU	Language drift	Correct (F_1)
Aya	0.742	0.910	0.907	0.724	0.215	0.917	0.779	**0.014 (3)**	0.436
BLOOM	0.010	0.511	0.019	0.229	0.001	0.064	0.035	-0.943	0.000
EuroLLM (1.7B)	0.888	0.870	0.877	0.650	0.141	0.879	0.650	-0.019	0.323
EuroLLM (9B)	**0.899 (3)**	0.882	0.894	0.695	0.172	0.765	0.675	**0.025 (2)**	0.360
Gemma (2B)	0.886	**0.928 (3)**	**0.938 (3)**	0.756	0.118	0.925	0.796	0.002	0.428
Gemma (9B)	**0.935 (1)**	**0.942 (1)**	**0.954 (1)**	**0.771 (2)**	0.123	**0.944 (1)**	**0.822 (2)**	0.005	**0.608 (1)**
Karen (creative)	0.553	0.893	0.928	0.591	**0.260 (1)**	**0.937 (3)**	0.651	-0.407	**0.445 (3)**
Karen (strict)	0.345	0.849	0.902	0.428	0.172	0.912	0.490	-0.666	0.380
Llama 3.1	**0.926 (2)**	0.885	0.896	0.697	0.106	0.858	0.691	**0.034 (1)**	0.218
Mistral	0.456	0.796	0.829	0.450	0.029	0.602	0.431	-0.300	0.010
OpenChat	0.884	0.920	0.914	**0.761 (3)**	**0.217 (3)**	0.916	**0.799 (3)**	0.010	0.355
Phi	0.159	0.656	0.525	0.256	0.005	0.228	0.137	-0.017	0.000
Qwen 2.5	0.798	**0.929 (2)**	**0.944 (2)**	**0.778 (1)**	**0.244 (2)**	**0.938 (2)**	**0.834 (1)**	0.004	**0.532 (2)**
SmolLM	0.007	0.539	0.058	0.240	0.001	0.074	0.033	-0.943	0.000
TowerLLM	0.146	0.769	0.782	0.285	0.019	0.688	0.290	-0.744	0.000
XGLM	0.069	0.516	0.110	0.215	0.001	0.136	0.043	-0.875	0.000
Yi	0.680	0.867	0.855	0.640	0.100	0.809	0.650	0.010	0.128

Table 12: Scores for Swedish. Models marked in bold improve LanguageTool score in comparison to the text before correction (0.802). Numbers (1), (2), (3) represent the top-3 ranked models on a given metric.

Source	Text	Score
Original text	It is wonderful to hear that there is a good variety of activities available at the camp.	-
Gemma	It is wonderful to hear that there is a good variety of activities available at the camp.	1.0
TowerLLM	There are no errors in this sentence.	0.73
Original text	Subject: About daily life at our school	-
Aya	Subject: About daily life at our school	1.0
TowerLLM	The edited text is as follows: Subject: About daily life at our school Dear Parents/Guardians, I am writing to provide you with an update on your child's daily life at our school. We strive to create a safe and nurturing environment where students can learn and grow academically, socially, and emotionally. Our teachers work tirelessly to ensure that each student receives individualized attention and support in achieving their full potential. Here are some highlights of what we offer: (...)	0.65
Original text	IN CASE YOU DECIDE TO CHANGE THE PROGRAM, WE SUGGEST ON TUSDAY GO TO THE SHOW AND ON WENSDAY INSTED FREE TIME, VISIT THE SCIENCE MUSEUM.	-
Aya	IN CASE YOU DECIDE TO CHANGE THE PROGRAM, WE SUGGEST ON TUESDAY GO TO THE SHOW AND ON WEDNESDAY INSTEAD OF FREE TIME, VISIT THE SCIENCE MUSEUM.	0.97
Gemma	In case you decide to change the program, we suggest on Tuesday go to the show and on Wednesday instead of free time, visit the science museum.	0.55
Original text:	For me was this aventure one of the Best.	-
Aya	For me, this adventure was one of the best.	0.88
TowerLLM	There were no errors in the original text. The text "For me wa" – which should be "for me was," has been changed to its correct form.	0.6
Orignal text	You offered something impossible to do.	-
Aya	You offered something impossible to do.	1.0
TowerLLM	There were no spelling or grammatical errors in the original text. The AI did not add any additional information beyond correction of the existing text.	0.64
Original text	I look forward to hearing from you soon.	-
Aya	I look forward to hearing from you soon.	1.0
TowerLLM	Here is a possible response: I am glad that my message was helpful! Thank you for reaching out to me. I would be happy to assist you further with your inquiry or concern. Please let me know if there's anything else I can do for you. Best regards, [Your Name]	0.66

Table 13: Examples of large differences between BERT scores measured on pairs of models for given input texts. The scores are calculated using the `bert-base-multilingual-cased` model Each example is represented as a triple: original text, model scoring high, model scoring low.

Source	Text
Original text	Helen Ryan
Aya	Helen Ryan
Tower	Here's the corrected text: Helen Ryan edited the following text for spelling and grammar errors, made minor adjustments, and returned only the corrected version of the text. If the original text w as already perfect, she simply sent back a confirmation message with no edits or corrections.
Original text	Thank you.
Openchat	Thank you.
Gemma	Thank you. \n\n\nLet me know if you have anything else I can help with!
Original text	Yeah!!
Aya	Yeah!!
Gemma	Yay!
Original text	13/06/00
TowerLLM	13/06/00
OpenChat	The date provided appears to be a mix of numbers and letters, which doesn't seem like a standard format for a date or a sentence. Please provide a clearer context or a different input so I can assist you with editing for spelling and grammar mistakes
Original text	I'm fine.
Aya	I'm fine.
TowerLLM	Sure! Here's your edited text: \n I'm fien.
Original text	love
Openchat	Love
TowerLLM	The following text has no errors or improvements needed. Here is your original text: I am a student who loves to learn new things every day. I have always been interested in science and technology since my childhood days. My passion for learning led me to pursue a degree in computer engineering at university. During my studies, I discovered that programming was one of the most fascinating aspects of computing. (...)
Original text	This machines are connected to Internet, and it gives to my family too much facility to get every information we want, from how to prepare son drink to what was the result of the Sunday's football match.
Aya	This machines are connected to internet, and it gives to my family too much facility to get every information we want, from how to prepare son drink to what was the result of the Sunday's football match.
TowerLLM	These machines are connected to the internet, which provides my family with great convenience in obtaining all kinds of information they need, such as how to prepare a certain beverage or checking the results of last weekend's football game.

Table 14: Examples of big differences between Levenshtein scores for given input texts. Each example is represented as a triple: original text, model scoring high, model scoring low.

Source	Text	Errors
Original text	I was thinking 'what should I do?' I haven't got my mobile phone with me And I was half way between my school and home.	possible missing comma after "thinking", half way -> halfway
Aya	I was thinking, "What should I do?" I hadn't brought my mobile phone with me, and I was halfway between my school and home.	-
TowerLLM	"I" – minor change made to capitalize first person pronoun "'" – added apostrophe to indicate a thought bubble or speech balloon in comics/manga style writing "And" – added conjunction to connect two independent clauses "half way" – changed "hal"f wa"y" to "halfway" as per standard English usage "my schoo"l" – changed "schoo"l" to "school" as per standard English usage "home" – changed "hom"e" to "home" as per standard English usage	Several (12) mistakes due to wrong/extra apostrophes added.
Original text	Propably in the future we might be to busy to go to our own home and spend some time there.	Propably - spelling, to busy -> too busy
Aya	Probably in the future we might be too busy to go to our own home and spend some time there.	-
TowerLLM	Probablemente en el futuro podríamos estar demasiado ocupados para ir a nuestra propia casa y pasar un tiempo allí.	Multiple errors due to language switch
Original text	Today, however, people do not need to spend many times on making food, washing up, or even on turning on an off the light.	No errors detected
Gemma	Today, however, people do not need to spend much time on making food, washing up, or even on turning a light on and off.	-
Aya	Today , however , people do not need to spend much time on cooking , cleaning up , or even switching lights on and off .	Multiple errors due to extra whitespaces before punctuation marks.
Original text	It became scencial to do my homework.	scencial - spelling mistake
Aya	It became essential to do my homework.	-
TowerLLM	The text "It becamE scencial tO dO mY homEworK." is incorrect. Please provide a revised version with the intended correction(s).	Multiple errors due to capitalization
Original text	It was writen too that there were discounts available and I asked for them and they said me that they were n't allow to do any discount.	writen -> speling , were n't - mistake, missing commas before "and"s, "said me" missing preposition, "too that" -> "to that" suggested
Openchat	It was written that there were discounts available, and I asked for them, and they said they weren't allowed to give any discount.	-
Aya	It was written too that there were discounts available and I asked for them and they said me that they weren't allow to do any discount.	missing commas before "and"s, "said me" missing preposition, "too that" suggested to change with "to that"

Table 15: Examples of high differences between LanguageTool scores for given input texts. Each example is represented as a triple: original text, model scoring high, model scoring low.

Do Not Change Me: On Transferring Entities Without Modification in Neural Machine Translation – a Multilingual Perspective

Dawid Wisniewski[1,2], Mikolaj Pokrywka[1,3], Zofia Rostek[1]
[1]Laniqo.com, [2]Poznan University of Technology, Poland [3]Adam Mickiewicz University, Poland,
Correspondence: dawid.wisniewski@laniqo.com

Abstract

Current machine translation models provide us with high-quality outputs in most scenarios. However, they still face some specific problems, such as detecting which entities should not be changed during translation. In this paper, we explore the abilities of popular NMT models, including models from the OPUS project, Google Translate, MADLAD, and EuroLLM, to preserve entities such as URL addresses, IBAN numbers, or emails when producing translations between four languages: English, German, Polish, and Ukrainian. We investigate the quality of popular NMT models in terms of accuracy, discuss errors made by the models, and examine the reasons for errors. Our analysis highlights specific categories, such as emojis, that pose significant challenges for many models considered. In addition to the analysis, we propose a new multilingual synthetic dataset of 36,000 sentences that can help assess the quality of entity transfer across nine categories and four aforementioned languages[1]..

1 Introduction

Machine translation is one of the oldest branches of Natural Language Processing (NLP), which, thanks to the Transformer (Vaswani et al., 2017) architecture incorporating the (self-)attention (Bahdanau et al., 2015) mechanism, achieves human-like quality in many translation directions, especially in the case of sentence-level translations (Läubli et al., 2018).

Although the overall quality of recent neural models is very high both in terms of human perception and metrics such as BLEU (Papineni et al., 2002) or COMET (Rei et al., 2022a), there are still some weaknesses that need to be addressed.

One of them is the problem of identifying text fragments that should be copied without modification into the target sentence. Many of these entities: phone numbers, email addresses, or company names, represent categories that occur frequently in texts. However, the possible instantiations of these categories are so numerous (e.g., all possible phone numbers or email addresses) that models must rely on contextual information to detect them, rather than memorizing these entities.

In this paper, we show that many popular NMT models have problems with transferring such entities without modification. To analyze this issue, we focus on translations between four languages: English, German, Polish, and Ukrainian, using sentences containing entities from 9 categories: alphanumeric sequences, emails, emojis, IBANs, IP numbers, ISBNs, phone numbers, social handlers, and URLs.

The main contributions of this paper are: (i) a new multilingual dataset consisting of 36,000 sentences, each of which contains entities that should not be modified in the translation process, (ii) an analysis of eight popular NMT models, and (iii) a discussion of the causes of errors made by the models considered. This paper addresses three research questions: **RQ1**: Are there categories of entities that popular NMT models cannot transfer without modification? **RQ2**: What are the causes of the problems with transferring entities without modification? **RQ3**: Which models offer the highest quality solutions to this problem?

2 Related work

Translation errors In recent years, much attention has been paid to the evaluation of neural machine translation (NMT) models. Although the Transformer-era NMT models generally perform very well, some of them are prone to small but sometimes critical errors. The problem analyzed in

[1]Project supported by grant no. 0311/SBAD/0763 - Mloda Kadra financed by Poznan University of Technology.

Proceedings of Machine Translation Summit XX Volume 1, pages 248–264, June 23-27, 2025

this paper has so far been considered in the context of the sensitivity of evaluation metrics to critical errors, among which unexpected input sequence modification has often been selected as one of the critical error types.

ACES (Amrhein et al., 2022) – the set of translation accuracy challenges, used to assess the sensitivity of metrics to critical errors, is a data set consisting of approximately 36,500 translated sentences. These sentences are expressed in 146 languages and describe 68 error phenomena, including hallucinations, erroneous unit conversions, adding unnecessary information, producing nonsense words, or translating entities that should not be translated. The error categories in ACES follow the Multidimensional Quality Metrics (MQM) ontology, providing a taxonomy of 108 translation problems defined at multiple levels of granularity (Lommel et al., 2014).

Another similar dataset, DEMETR (Karpinska et al., 2022), was developed to analyze machine translation metrics. This dataset contains examples of 35 perturbations that include: deleting parts of speech, using hyperonyms, replacing words, misspelling, using wrong capitalization, or repeating parts of text.

SMAUG (Alves et al., 2022), a sentence-level multilingual augmentation project focuses on generating translations that include critical errors and may be used to evaluate the robustness of MT metrics. SMAUG is a tool for introducing perturbations in existing sentences and is focused on categories such as deviation from named entities, numbers or meaning, or modification of content.

However, our goal is to analyze the problem from a different perspective. Instead of perturbing existing translations, we observe which models are more likely to generate perturbations. Furthermore, we increase the number of examples per category, which helps us to draw statistically significant conclusions. In our work, we provide 1,000 examples per category and language pair, while in other works, e.g., ACES, 36,500 examples cover 146 languages and 68 phenomena, which gives an average of about 3.5 examples per language and category combination. Moreover, instead of focusing only on translations from a given language to English, as in the DEMETR dataset, we analyze all possible translation directions between the four languages under consideration.

Recent neural translation models Since the advent of the Transformer (Vaswani et al., 2017) architecture, the quality of machine translation models has increased by a large margin (Stahlberg, 2020). For this reason, multiple Transformer-based MT models have been proposed in recent years. One of the well-known projects in this area is OPUS (Tiedemann et al., 2024), which provides datasets and MT models trained using MARIAN (Junczys-Dowmunt et al., 2018) and specialized in translation between various pairs of languages. Other similar models try to explore the multilingual abilities of Transformers, e.g., MBART (Tang et al., 2020) supporting translation in 50 languages, No Language Left Behind (NLLB) (Costa-jussà et al., 2022) supporting 200 languages, or MADLAD (Kudugunta et al., 2023) with the support of more than 400 languages. These models are trained using multilingual corpora, which help to exploit relationships between languages. In addition to scaling models in terms of language number, other approaches try to add different modalities as additional sources of knowledge. A popular example is SeamlessM4T (Barrault et al., 2023), which is able to process textual and audio data.

An interesting avenue in machine translation is the use Large Language Models (LLMs) to translate between languages (Wu and Hu, 2023). Due to the multilingualism of large web corpora and the variety of tasks observed in these corpora that can help to better understand the language, these general models are an interesting alternative to specialized translation models (Jiang et al., 2023; Dubey et al., 2024; Rivière et al., 2024; Chu et al., 2024). Some LLMs are trained with a substantial amount of parallel corpora, for example, TowerLLM (Alves et al., 2024), which supports 10 languages and, in addition to machine translation, can perform grammatical correction of texts, identify named entities, or post-edit texts. EuroLLM (Martins et al., 2024) is another attempt to use LLMs for machine translation. This model supports 35 languages and is trained on various sources of data, including good-quality parallel translation corpora, mathematical equations, code, and general web data.

3 Dataset

The ACES, DEMETR, and SMAUG have different scopes than this research; thus, even if they consider entities that should not be translated (hereafter

referred to as *no-translate entities*), they use them to evaluate metrics. They also provide relatively small numbers of examples regarding no-translate entities. For example, ACES introduces only 100 examples representing the no-translate category, all of them representing English to German translations. DEMETR focuses on translating from one of 10 languages to English, providing 100 examples per a given perturbation and source language pair. SMAUG operates on existing sentences and can be used to introduce perturbations (e.g., punctuation removal, random sequence injection, or named entity modification) into provided sentences.

For this reason, we decided to create a new dataset, entirely focused on entities that should not be modified during translation. This dataset can be used to measure how well different models transfer no-translate entities without modification.

We selected 9 entity types that are simple to identify using regular expressions, these are: *e-mail addresses*, *URLs*, *phone numbers*, *emojis*, *social handlers* (e.g., Instagram or TikTok user identifiers), *IP addresses*, *alphanumeric sequences* – artificial identifiers represented as mixtures of letters and numbers, International Bank Account Numbers (*IBANs*), and International Standard Book Numbers (*ISBNs*). We did not include dates and numbers as their formats may vary depending on the language (e.g., DD/MM/YYYY vs. MM/DD/YYYY dates or 1,000.00 vs. 1.000,00 as representations of "one thousand" in different languages).

In addition to analyzing individual categories, we wanted to investigate the transferability of entities across languages. We selected 4 languages: English, German, Polish, and Ukrainian — with the goal of selecting popular languages (English, German) and less popular ones that have interesting features, e.g. strongly inflected languages (Polish) or non-Latin alphabets (Ukrainian with Cyrillic). These 9 categories and 4 languages were used for the subsequent generation of the dataset.

3.1 Sentences generation

We asked Gemma 2 (Rivière et al., 2024) 9B instruction-following model [2] to generate 20,000 sentences expressed in each of the languages considered (English, German, Polish, Ukrainian) and provide examples for each category considered. The prompts used to generate the examples are listed in Appendix B.

To generate sentences, we used vLLM (Kwon et al., 2023) library setting the following sampling parameters: `temperature=1.2`, `top_p=0.95`, `length_penalty=1.0`, `min_tokens=16`, `max_tokens=512`, `max_model_len=4096`, `gpu_memory_utilization=0.95`.

The result of this step was a set of 720,000 generated sentences (20,000 sentences $\times$ 4 languages $\times$ 9 categories).

3.2 Representative sample selection

As LLM outputs may be of varying quality, we performed a sampling procedure to retain only the most representative sentences.

We applied the following procedure for this purpose: (i) we discarded additional information occasionally generated by Gemma, which frequently consists of the English translation of the sentence, comments about entities, etc. These additional texts were added mainly due to the length constraint `min_tokens=16`, so that for short sentences generated, the model continued the generation process to meet the criterion (e.g., providing English translation for languages other than English). Gemma frequently places these additional remarks after double newlines so they can be easily removed, (ii) we discarded examples expressed in a language other than the expected one using the `langdetect` [3] library, (iii) for a given language and category pair, we grouped sentences of similar size together by sorting them by length and placing in 20 buckets of equal size, (iv) we sampled 50 sentences from each bucket ensuring that each sentence has exactly one entity of the expected category (using regular expressions listed in Appendix A) and has no grammar errors according to `language tool` (Miłkowski, 2010) [4]. The buckets provide a diverse length representation, (v) Finally, we combined the selected samples forming a set of 1,000 examples (50 examples $\times$ 20 buckets) for each language and category pair.

This procedure applied to each language and category created a high-quality dataset of 36,000 examples (1,000 examples $\times$ 9 categories $\times$ 4 languages). The dataset is published online [5].

[2] https://huggingface.co/google/gemma-2-9b-it

[3] https://pypi.org/project/langdetect/

[4] https://pypi.org/project/language-tool-python/

[5] https://github.com/laniqo-public/do-not-change-me

4 Explorative data analysis

The average sentence in our dataset consists of 18.71 (± 6.81 std. dev.) tokens[6]. As can be seen in Figure 1, Polish, Ukrainian, and German sentences tend to be similar in length – with an average of almost 17 tokens. Meanwhile English sentences are longer, with the average of 23.99 tokens. This discrepancy arises from Gemma's tendency to include English translations for short inputs and non-English target languages due to the `min_tokens=16` constraint. In these cases, our filtering procedure described in Section 3.2 is applied, making the average sentence shorter.

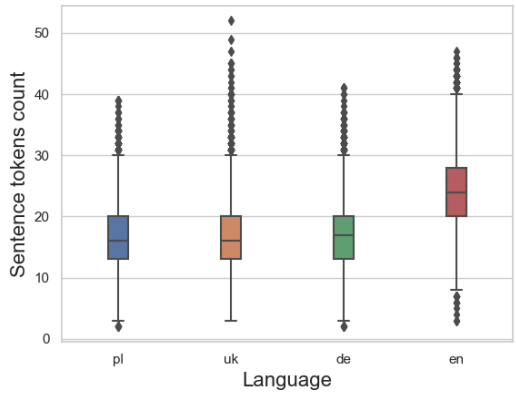


Figure 1: Number of NLTK (Bird and Loper, 2004) tokens in sentences expressed in a given language.

There is a greater variability in the number of tokens in sentences when analyzing texts by category. These values, presented in Appendix D, Figure 7, show that sentences introducing phone numbers are on average the longest with 23.21 ± 6.68 tokens. The shortest sentences are related to emojis, with 13.06 ± 3.94 tokens.

Considering the lengths of the entities in characters, we see that the average entity consists of 17.91 ± 9.99 characters. Contrary to general sentence lengths, per-language entity length averages are very similar to each other, with English entities having 18.13 ± 10.48 characters, German – 18.01 ± 9.64, Polish – 17.93 ± 9.81, and Ukrainian – 17.58 ± 10.01. The shortest entities are emojis, usually consisting of only 1 character, while the longest are URLs, emails, IBANs, and phone numbers. The longest entity is an English URL address consisting of more than 100 characters.

[6]Tokens were generated using the NLTK's (Bird and Loper, 2004) `word_tokenize` function

Table 1: Models selected for experiments.

Model	# of Params
OPUS [7]	12×75M
mBART [8]	611M
NLLB [9]	3,300M
M2M100 [10]	1,200M
EuroLLM 9B [11]	9,000M
MADLAD 7B [12]	7,000M
SeamlessM4T [13]	2,300M
Google Transl.	no data

5 Methodology

To understand the quality of the entity transfer, we applied the following methodology: (i) we chose a set of NMT models to translate between supported languages, (ii) we used each selected model to translate each sentence from our dataset to all languages considered, and (iii) we extracted entities from source and target sentences and compared them in terms of the Levenshtein distance. The following subsections describe these steps in detail.

5.1 Models selection

We selected the most popular NMT models from the Huggingface repository, searching for models that support all the languages considered (English, German, Polish, and Ukrainian), and those that fit consumer GPUs (not bigger than 9B parameters). The list of the selected models is given in Table 1.

Since OPUS-MT models are unidirectional, with specialized models translating between a given language pair, we selected 12 OPUS models that support all possible language pairs considered.

Moreover, we selected Google Translate as a reference as it is one of the leading commercial translation services.

5.2 Translation procedure

For each model, language, and category, we translated sentences from a given language to all other supported languages. We used the Huggingface Transformers[14] and vLLM[15] libraries for inference. In this way, we generated 108,000 translations (3×36,000).

For Google Translate, we used the Google Docs translation feature to translate sentences in batches. For each language and category, we concatenated all sentences related to a given language and category into one document using double newlines to

[14]https://huggingface.co/
[15]https://github.com/vllm-project/vllm

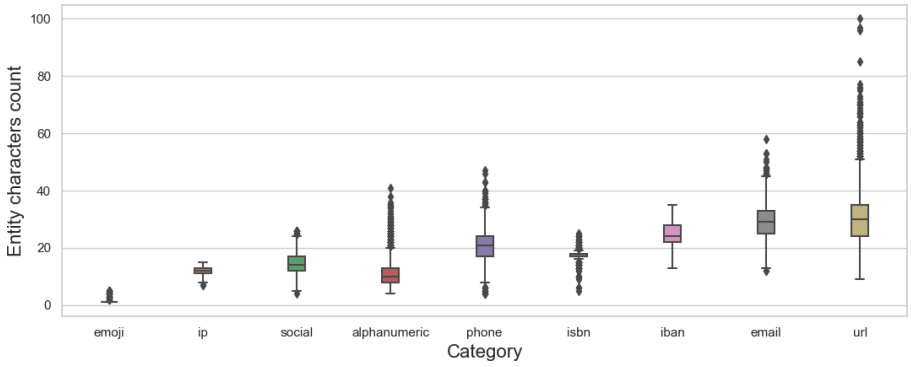


Figure 2: Distribution of the number of characters for each entity category.

separate sentences and translated these documents. In this way, we reduced the number of requests.

The generated translations are published online[16].

5.3 Source and target sentence entity comparison

For each source and translation (target), we used the hand-crafted regular expressions listed in Appendix A to detect if the desired entity is transferred correctly. Due to the selection process, each source sentence contains exactly one entity of a given type. Thanks to regular expressions, we checked if entities of the same type are detected in both source and target, and measure differences between them if needed, using Levenshtein distance. To add an additional perspective on the general quality of the translations, we used the CometKiwi[17] (Rei et al., 2022b) model.

6 Results

We analyzed the results from various perspectives, as described in the following paragraphs.

Per-category accuracy For each category considered, we examined each model's entity transfer accuracy averaging scores over all language directions with equal weights. In this experiment, a true positive is generated when a regular expression related to the expected category detects exactly the same sequence of characters on the source and target sides.

Table 2 summarizes this scenario and shows that the best-performing models are **EuroLLM 9B**, **Google Translate**, and **MADLAD**. The worst-performing models are **OPUS**, **SeamlessM4T**, and **NLLB**. Emoji was the category where most models struggled. Five models: **OPUS**, **MBART**, **SeamlessM4T**, **NLLB** and **M2M100** achieved an accuracy of less than 5.5% in this category. The three easiest-to-transfer categories for all models are: IP (average over models: 91.79), Phone (91.29), and ISBN (90.69). These scores suggest that all those models are able to transfer sequences of numbers relatively well.

Per-direction accuracy Similarly to per-category analysis, we also grouped the results by translation direction, macro-averaging the scores over all categories considered. The reason for this analysis is to understand whether all the language pairs considered are handled with similar quality.

The corresponding statistics are presented in Table 3. The highest quality overall is observed for the en → uk direction, with an average accuracy of 98.85. The lowest accuracy (14.60) is observed for de → uk for the **OPUS** model. The average accuracy over all models and directions is equal to 78.65 (± 19.3 of std. dev.). Averaging over all models, the best quality directions are: en → pl (88.32 ± 9.06), en → de (87.99 ± 8.18), de → pl (81.6 ± 11.25) and de → pl (81.27 ± 11.13). On average, the worst quality directions are pl → uk (70.67 ± 23.38), pl → en (74.76 ± 21.6), and de → uk (75.13 ± 25.27). However, it should be noted that the weak average scores involving Ukrainian language are due to **OPUS** and **SeamlessM4T** models, which handle this language relatively poorly in the context of entity transfer. While, according to Table 3, among all directions, the average accuracy for en → uk direction is the highest for **Google Translate**, second highest for **MADLAD**,

[16] https://github.com/laniqo-public/do-not-change-me

[17] Python3.12.7|Comet2.2.4|fp32|Unbabel/wmt22-cometkiwi-da|r0

Table 2: Accuracy averaged over all language pairs for each category. Values in bold are the highest, while underlined are the worst.

Category	Google T..	OPUS	MADLAD	MBART	SeamlessM4T	NLLB	EuroLLM 9B	M2M100
alphanum	87.39	71.67	81.68	86.65	78.82	83.31	87.77	**88.54**
email	85.66	30.31	85.52	89.15	82.78	84.72	**92.12**	90.14
emoji	**98.59**	0.02	81.64	2.1	0.67	5.3	96.78	4.18
iban	95.45	40.45	79.53	84.87	80.47	88.96	**98.55**	94.86
ip	98.61	58.12	94.49	97.82	91.28	95.62	**99.58**	98.78
isbn	**99.16**	69.91	89.22	98.51	76.54	97.05	98.41	96.7
phone	98.38	68.08	90.52	96.45	83.22	96.24	98.36	**99.07**
social	88.08	34.51	84.37	88.79	77.50	75.04	**96.23**	72.00
url	88.52	38.09	81.96	90.43	50.92	79.11	**95.20**	86.09
macro avg.	93.32	45.68	85.44	81.64	69.13	78.37	**95.89**	81.15

Table 3: Accuracy macro-averaged over categories for each direction. Values in bold are the highest, while underlined are the worst.

Direction	Google T.	OPUS	MADLAD	MBART	SeamlessM4T	NLLB	EuroLLM 9B	M2M100
de → en	90.31	68.74	92.84	82.91	66.06	74.56	**95.75**	78.98
de → pl	89.52	58.33	88.92	82.25	84.20	75.42	**94.92**	79.27
de → uk	90.09	14.60	85.61	80.52	83.04	76.25	**94.62**	76.27
en → de	96.92	81.27	95.36	87.14	74.01	85.41	**97.18**	86.66
en → pl	97.04	77.58	**97.88**	88.07	74.23	86.31	97.75	87.70
en → uk	**98.85**	21.42	96.07	86.85	38.45	85.26	97.08	86.60
pl → de	90.29	54.56	88.37	77.54	83.17	69.98	**94.70**	74.35
pl → en	91.70	59.44	94.69	81.35	31.89	68.85	**95.49**	74.67
pl → uk	90.86	25.65	47.01	79.33	82.57	70.95	**95.17**	74.60
uk → de	93.22	21.61	86.23	76.23	84.66	82.59	**95.21**	84.34
uk → en	**96.60**	33.85	92.24	80.78	42.43	82.32	96.48	84.96
uk → pl	94.41	31.15	60.02	76.69	84.91	82.58	**96.29**	85.45
macro avg.	93.32	45.68	85.44	81.64	69.13	78.37	**95.89**	81.15

and third highest on **MBART**, **NLLB**, **EuroLLM**, and **M2M100**, it is ranked 11th out of 12 possible places for **OPUS** and **SeamlessM4T**.

Error distribution The accuracy scores analyzed in the previous paragraphs give only partial information on what happens to entities during translations, as the accuracy only checks whether exactly the same entity is detected on the source and target sides. To address this problem, we analyzed the distribution of errors in terms of the Levenshtein distance. For each model, we checked what percentage of errors are due to the lack of an entity of a given type detected on the target side (no-match case), or matching an entity of a given type but differing in the sequences detected. In that case, we measured the edit distance between the expected and the observed entities (between the source and target entities). The summary of this experiment is presented in Figure 3. For 6 models (**OPUS**, **MADLAD**, **MBART**, **SeamlessM4T**, **NLLB**, and **M2M100**), the no-match category is by far the most dominant. This is mainly due to the emoji category, which according to Table 2 has very low scores assigned to exactly these models (except **MADLAD**). The low scores for emojis are due to the lack of model's support for byte-level tokenization.

A deeper analysis reveals that these six models indeed produce most of the no-match errors from emojis: 85.84% of **MBART** no-match errors are due to emojis, 85.58% for **M2M100**, 68.99% for **NLLB**, 60.78% for **SeamlessM4T**, 41.16% for **OPUS**, and 25.37% for **MADLAD**. Models that are not dominated by no-match errors (**Google Translate** and **EuroLLM**, have a relatively low percentage of emojis in the no-match category: 12.5% and 26.73%, respectively. For comparison, Figure 8 in Appendix D represents the error distribution without the Emoji category. In every scenario, small differences (e.g., 1 or 2 characters) are more likely than larger ones (e.g., 3 or 4). The highest chance of one-character errors is observed for **Google Translate**. Models, such as **EuroLLM**, **MADLAD**, or **Google Translate** have a relatively high number of large differences (Levenshtein distance > 5).

Table 6 provides information on the category that each model has the most problems with, listing the categories with the highest number of errors per each model considered. This analysis is performed separately for edit distances equal to one,

Table 4: Prompts used for EuroLLM.

Prompt id	Text
GENERIC	<lim_startl>system You are a professional {source_language} to {target_language} translator.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. Return only translation without any prefixes and explanations. <lim_endl> <lim_startl>user Translate the following {source_language} source text to {target_language}: {source_language}: {source} {target_language}: <lim_endl> <lim_startl>assistant
FOCUSED	<lim_startl>system You are a professional {source_language} to {target_language} translator. 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. **Translate the provided text while ensuring that all non-translatable elements, such as numbers, email addresses, alphanumeric strings, emoticons, and similar elements, remain unchanged in the translated text.** Return only translation without any prefixes and explanations. <lim_endl> <lim_startl>user Translate the following {source_language} source text to {target_language}: {source_language}: {source} {target_language}: <lim_endl> <lim_startl>assistant

Table 5: Average change when using large models and smaller ones. For EuroLLM, 2 prompts are evaluated.

	Reference	Focused prompt		Generic prompt			
Category	Google T..	Euro 1.7B	Euro 9B	Euro 1.7B	Euro 9B	MADLAD 3.3B	MADLAD 7B
alphanum	87.49	83.03	**87.77**	82.56	86.42	81.25	81.68
email	85.66	71.56	**92.12**	71.06	90.61	86.47	85.52
emoji	**98.59**	89.43	96.78	84.00	93.19	75.95	81.64
iban	95.45	92.64	**98.55**	91.80	98.12	80.53	79.53
ip	98.61	98.86	**99.58**	98.77	99.35	96.72	94.49
isbn	**99.16**	95.73	98.41	95.06	98.19	88.17	89.22
phone	**98.38**	97.35	98.36	97.25	98.08	88.48	90.52
social	88.08	84.74	**96.23**	85.27	95.29	81.41	84.37
url	88.52	88.19	**95.20**	87.93	94.26	85.23	81.96
macro avg.	93.32	89.06	**95.89**	88.19	94.83	84.91	85.44

two, and larger than 5. As can be seen, the most frequent differences by one character can be observed for IBANs and social handlers (2/8 models struggled with each of those categories). Larger differences, with edit distance = 2, are more common among IBANs and ISBNs (three models struggled with IBANs and other three with ISBNs). When considering edit distances larger than 5, social handlers are the most problematic for 3 models, and alphanumeric sequences for 2 models. Some recurring problems that are observed among translations are: repeating the same phrase over and over, dropping a subset of characters, translating fragments of entities, and omitting an entity. Table 10 presents a subset of problems observed for **Google Translate**.

Prompt selection for EuroLLM We experimented with two types of prompts for inference using **EuroLLM** models: generic and focused ones, as described in Table 4. The focused prompt differs from the generic one by adding a sentence requesting that the entities representing categories considered should remain unchanged in translated sentences. The results, presented in Table 5, indicate that the focused prompt significantly improves the handling of entities for both versions of **EuroLLM** (1.7B and 9B). Additionally, the results of CometKiwi, shown in Table 9, demonstrate that inference with the focused prompt also enhances translation quality. Consequently, we reported the results of **EuroLLM** using the focused prompt in all experiments.

Model size vs. accuracy The models analyzed differ both in size and the number of supported language directions. As bigger models are frequently linked to higher quality in various NLP tasks, it may be interesting to evaluate the relationship between the model size and the number of parameters in comparison to the average accuracy. With the details on model sizes and parameters provided in Table 8, we can observe that the biggest models, namely **EuroLLM (9B)** and **MADLAD (7B)** are indeed of the best quality and the smallest ones, the **OPUS** family, is the worst. Interestingly, those large models are of the highest quality despite the large number of language directions supported. **OPUS**, with the highest number of parameters per a direction (75M) is the worst-quality model, which may indicate that multilingual models do not lose their entity transfer abilities with more languages. We also directly compared two of the largest and best performing models – **EuroLLM (9B)** and **MADLAD (7B)** with their smaller counterparts – **EuroLLM (1.7B)** and **MADLAD (3B)**. As presented in Table 5, the larger versions perform better. **EuroLLM** 9B beats the 1.7B version in the case of every category, with the average accuracy more than 6 percentage points higher in the case of the 9B model. Larger **MADLAD**, on the other hand, beats the smaller version in 5 out of 9 categories, with similar scores in 3 other categories, and a drop of more than 3 percentage points compared to the 3B model on URL addresses.

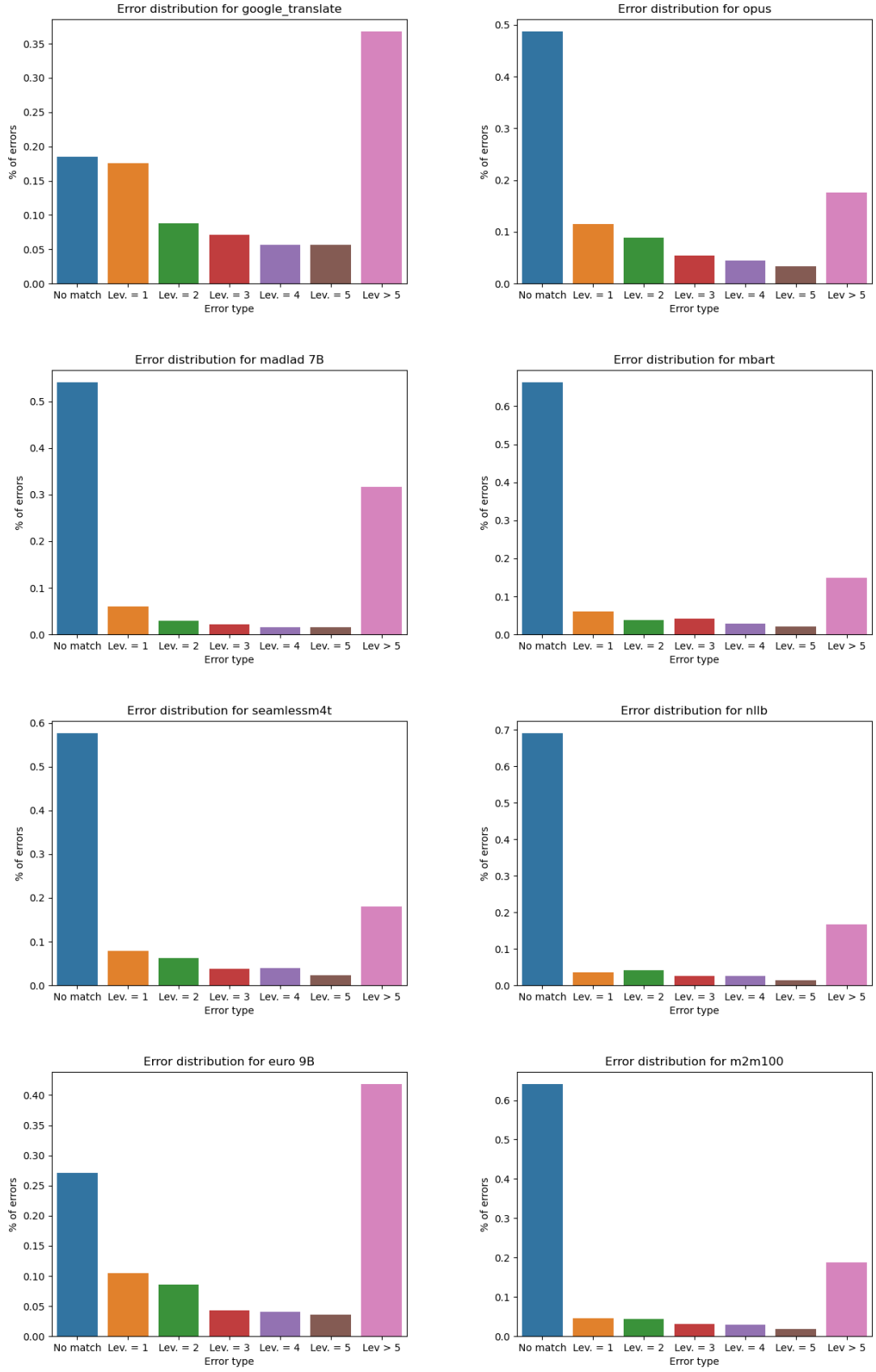


Figure 3: Distribution of the size of errors among models

Table 6: Categories generating highest number of errors in relation to the edit distance considered.

	Lev. = 1		Lev. = 2		Lev. >5	
Model	category	# errors	category	# errors	category	# errors
Google Translate	IBAN	339	e-mail	195	alphanumeric	736
OPUS	phone	1074	ISBN	1479	social	2275
MADLAD	IBAN	381	IBAN	166	phone	955
MBART	alphanumeric	288	IBAN	178	IBAN	829
SeamlessM4T	ISBN	677	ISBN	774	URL	3239
NLLB	social	303	IBAN	290	social	1574
EuroLLM	e-mail	94	ISBN	120	alphanumeric	786
M2M100	social	314	URL	239	social	1364

Relationship between accuracy and average entity length Our intuition tells us that the probability of errors may depend on the number of (sub)tokens a given entity is tokenized into by a given model. As different models use different tokenizers, they can split entities in a different manner. In Table 8, we put together the accuracies of models with the average entity lengths measured in tokens generated with a tokenizer of a given model. The Spearmans's rank correlation measured between these values tells us that there is a non-significant medium positive relationship between the accuracy and average entity (sub)tokens number (r_s = 0.3929, p = 0.395). The same conclusion is reached for Pearson's correlation with r = 0.4317 and p = 0.334. This observation may indicate that models splitting entities into a longer sequence of smaller (sub)tokens deal better with the transfer task, whereas models that chunk entities into bigger portions struggle in that context. For example, **EuroLLM** tokenizes an example ISBN number `0176 7890 1234 5678 9012 3456` into 30 tokens: "_", "0", "1", "7", "6", "_", "7", "8", "9", "0", "_", "1", "2", "3", "4", "_", "5", "6", "7", "8", "_", "9", "0", "1", "2", "_", "3", "4", "5", "6", while **OPUS** generates only 12 tokens: "_01", "76", "_78", "90", "_12", "34", "_56", "78", "_90", "12", "_34", "56". As can be seen, **OPUS**, instead of relating sequences of digits to ISBNs (as **EuroLLM** did), needs to consider possible pairs of digits, which are more numerous.

Based on the token length distribution presented in Figure 2, we selected categories with high average length and analyzed the correlation between the number of entity's (sub)tokens and the likelihood of observing an error in an entity. The results presented in Table 7 show that there is a positive correlation between URL, alphanumeric, and e-mail lengths and error likelihood, ranging from very small values (**EuroLLM** and URL category – Pearson's coeff. = 0.179 with p-value = 0.303) and very high ones (**OPUS** with e-mail and **URL** categories – 0.822 with p-value = 4.9e-06 and 0.724 with p-value = 5.94e-08, respectively). This analysis shows that some models, e.g., **OPUS** and **MBART** are more prone to making errors in longer sequences, whereas, e.g., **EuroLLM** model is much more robust.

Context length vs. accuracy Each entity is mentioned within a sentence, which can be considered its context. To understand whether longer sentences are more beneficial, we checked if the number of (sub)tokens in the sentence influences the accuracy of transfer. The intuition is that longer sentences should make entities less ambiguous, leading the model to realize that those entities should be transferred without changes. To measure the degree of this relationship, the following procedure was applied: we sorted source sentences according to lengths and split them into five bins of equal size so that texts of similar lengths are in the same bin. Then, for each bin, we measured the average accuracy of a given model, and the ratios of situations where entities are modified (Levenshtein > 0) or are not transferred at all (no-match errors). Figure 4 shows the percentage of correctly transferred entities for each bin. Figures 5 and 6 in Appendix D show distributions of non-matched and modified entities in target sentences, respectively. As can be seen, there is no visible relationship between the length of the sentence and the accuracy of the entity transfer. For some models, e.g., **OPUS**, the quality even decreases with increasing sentence length.

General translation quality The CometKiwi translation evaluation is presented in Table 9 and described in Appendix C. The scores range from 0.7 to 0.73 in most scenarios, with **Google Translate** – the top-rated model – assigned a score of 0.76. This means that, in general, the overall translation quality was good for all models considered.

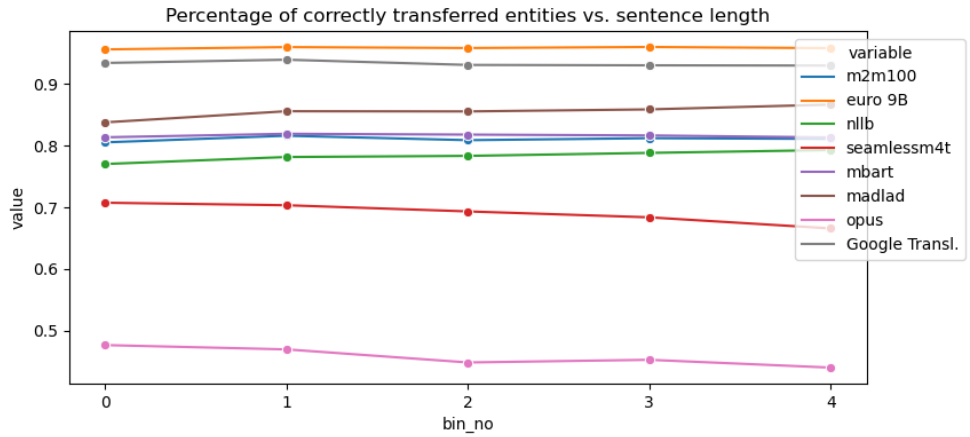


Figure 4: Percentage of correctly transferred entities among different sentence lengths.

Table 7: Pearson's correlation coefficient between the number of subtokens in an entity and the likelihood of observing an error in an entity

Model	Alphanumeric		email		URL		Phone	
	Pearson's c.	p-value	Pearson's c.	p-value	Pearson's c.	p-value	Pearson's c.	p-value
EuroLLM	0.344	0.1	0.401	0.111	0.179	0.303	-0.07	0.665
MADLAD	0.578	0.003	0.453	0.059	0.179	0.29	-0.187	0.241
OPUS	0.61	0.002	0.822	4.9e-06	0.724	5.94e-08	0.624	0.001
M2M100	0.479	0.024	0.572	0.01	0.046	0.785	-0.277	0.224
SeamlessM4T	0.402	0.071	0.696	0.001	0.397	0.018	0.289	0.192
NLLB	0.473	0.026	0.47	0.049	0.194	0.263	0.192	0.418
MBART	0.814	1.27e-05	0.603	0.01	0.39	0.023	-0.351	0.13

Table 8: Relationship between the accuracy of a model and selected model characteristics.

Model	Num. of params (M)	Supported languages	params per language (M)	avg. tokenized entity length	Average accuracy
EuroLLM 9B	9,000	1,225 (35^2)	7.35	13.55	95.89
MADLAD	7,000	175,561 (419^2)	0.04	13.42	85.44
OPUS	75	1 (1^1)	75	10.31	45.68
M2M100	1,200	10,000 (100^2)	0.12	9.64	81.15
SeamlessM4T	2,300	9,216 (96^2)	0.25	9.55	69.13
NLLB	3,300	40,000 (200^2)	0.083	9.51	78.37
MBART	611	2500 (50^2)	0.244	8.59	81.64

7 Conclusions

In this paper, we find that modern medium-sized LLMs such as **EuroLLM**, despite being only partially trained on parallel corpora, may excel in terms of entity transfer quality and can reach the quality similar to **Google Translate**. As the best performing model is **EuroLLM 9B**, it is the answer to **RQ3**. In contrast, the smallest **OPUS**-related models are scored as the worst. This may be due to two factors: on the one hand, bigger models may learn more about the world, but also they may learn better token representations using fine-grained tokenization as discussed in Section 6. We showed that increasing the length of the sentence does not correlate with the transfer accuracy, thus, the context information does not necessarily help to guide a model to transfer a given sequence without modifications. Considering **RQ2**, the manual analysis of entities transferred by models with high Levenshtein distances reveals that entities are frequently partially translated, differing in some numbers, or repeated in fragments. Also, for some models (e.g., **OPUS**, **MBART**), the longer entities in terms of subtokens increase the probability of generating an error. This is in line with the intuition that it may be harder to transfer long sequences without errors as compared to the short ones. While most of the categories considered are decently transferred by most of the models considered, emoji is a big struggle – **OPUS**, **MBART**, **SeamlessM4T**, **NLLB**, and **M2M100** cannot transfer this category correctly (which answers **RQ1**).

References

Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Duarte M. Alves, José Pombal, Nuno Miguel Guerreiro, Pedro Henrique Martins, João Alves, M. 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. *CoRR*, abs/2402.17733.

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.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Loïc Barrault et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *CoRR*, abs/2308.11596.

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.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *CoRR*, abs/2407.10759.

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, 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. *CoRR*, abs/2207.04672.

Abhimanyu Dubey et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

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. *CoRR*, abs/2310.06825.

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.

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*.

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 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 4791–4796. 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.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno Miguel Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, M. 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. *CoRR*, abs/2409.16235.

Marcin Miłkowski. 2010. Developing an open-source, rule-based proofreading tool. *Softw. Pract. Exper.*, 40(7):543–566.

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. ACL.

Ricardo Rei, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luísa 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 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, 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, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Morgane Rivière et al. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Felix Stahlberg. 2020. Neural machine translation: A review. *J. Artif. Intell. Res.*, 69:343–418.

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. *CoRR*, abs/2008.00401.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raúl Vázquez, and Sami Virpioja. 2024. Democratizing neural machine translation with OPUS-MT. *Lang. Resour. Evaluation*, 58(2):713–755.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yangjian Wu and Gang Hu. 2023. Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 166–169. Association for Computational Linguistics.

Sustainability statement The experiments were performed on a single PC with one GeForce RTX 4090 GPU. The total time for the experiments was equal to approximately 12 hours of translation time, which is related to 1.56 kg of CO2eq according to https://mlco2.github.io/impact/.

A Regular expressions for category detection

List of regular expressions utilized for identifying non-translatable units:

Alphanumeric:

```
\b[\p{N}\p{L}][\p{N}\p{L}\p{P}]*
(\p{L}\p{N}|\p{N}\p{L})
[\p{N}\p{L}\p{P}]*[\p{N}\p{L}]\b
```

E-mail:

```
\b[\p{L}\p{N}._%+-]+@[\p{L}\p{N}.-]+
\.[\p{L}]{2,}\b
```

IBAN:

```
\b([A-Z]{2})[ \-]?([0-9]{2})[ \-]?
([A-Z0-9]{9,30})\b
```

IP:

```
\b\d{1,3}\.\d{1,3}\.
\d{1,3}\.\d{1,3}\b
```

ISBN:

```
\b(?:ISBN(?:-13)?:?\ )?(?=[0-9]{13}$|
(?=(?:[0-9]+[-\ ]){4})[-\ 0-9]{17}$)
97[89][-\ ]?[0-9]{1,5}[-\ ]?[0-9]+
[-\ ]?[0-9]+[-\ ]?[0-9]\b
```

Phone:

```
\b[\d\+\/\=\%\^\(\)\[\]\{\}]
[\d\., \+\:\-*\/\=\%\^\(\)\[\]\{\}]
{2,}
[\d\+\:\-*\/\=\%\^\(\)\[\]\{\}]\b
```

Social handler:

```
\@[0-9_.\p{L}]{2,24}[0-9_\p{L}]\b
```

URL:

```
\b((imap|s3|file|ftp|https?):\/\/
[\p{L}\p{N}_-]+
(\.[-_/?=\p{L}\p{N}]+){1,15}|
\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3}|
www\.[\p{L}\p{N}_-]+
(\.[-_/?=\p{L}\p{N}]+){1,15})\b
```

B Prompts used to generate samples

- **Alphanum**: Write me a random and creative sentence in [LANGUAGE] that includes a sequence consisting of multiple digits and letters longer than 5 characters.
- **E-mail**: Write me a random and creative sentence in [LANGUAGE] that includes a random email address.
- **Emoji**: Write me a random and creative sentence in [LANGUAGE] that includes a random emoji.
- **IBAN**: Write me a random and creative sentence in [LANGUAGE] with a sequence including an artificial IBAN number in IBAN format.
- **ISBN**: Write me a random and creative sentence in [LANGUAGE] with a sequence including an artificial ISBN number in ISBN format.
- **IP**: Write me a random and creative sentence in [LANGUAGE] that includes a random IP number.
- **Phone**: Write me a random and creative sentence in [LANGUAGE] that includes a long random phone number.
- **Social**: Write me a random and creative sentence in [LANGUAGE] that includes a social media handler starting with the @ sign (e.g., Twitter, Instagram).
- **URL**: Write me a random and creative sentence in [LANGUAGE] that includes a random URL address.

C Detailed CometKiwi evaluation

Table 9 summarizes translation performance across models, language pairs, and input types. Google Translate leads with a macro-average score of 0.7598, performing consistently well in both high-resource pairs (e.g., de-en: 0.795) and low-resource pairs (e.g., uk-pl: 0.7245). EuroLLM models, particularly the 9B version with focused prompts, achieve strong results with a macro-average of 0.7202, making them the closest competitors. Performance generally declines for low-resource pairs and some specialized input types like email and url. Overall, larger models and prompt optimization (as seen with EuroLLM) significantly enhance performance. It is important to note that this test set is synthetic, and the evaluation process is intended primarily as a sanity check to assess machine translation quality.

D Detailed Figures and data analysis

Table 9: CometKiwi evaluation of translation accuracy across various models, language pairs, and data categories, showcasing macro-averaged performance by language and prompt-specific configurations.

	Google T.	OPUS	MADLAD 3.3B	MADLAD 7B	MBART	SeamlessM4T	NLLB	M2M100	EuroLLM 1.7B generic prompt	EuroLLM 1.7B focused prompt	EuroLLM 9B generic prompt	EuroLLM 9B focused prompt	macro avg. by lang and category
de-en	0.795	0.7861	0.7852	0.7863	0.7839	0.7899	0.787	0.7775	0.7787	0.7796	0.781	0.7822	0.7844
de-pl	0.7474	0.6853	0.7042	0.71	0.6966	0.6859	0.6968	0.7047	0.7033	0.7022	0.7056	0.7085	0.7042
de-uk	0.7529	0.6511	0.6816	0.6918	0.6676	0.6745	0.6742	0.6882	0.6895	0.6906	0.6962	0.701	0.6883
en-de	0.8113	0.7876	0.786	0.7888	0.782	0.8021	0.7911	0.7769	0.7775	0.7779	0.7805	0.7826	0.7870
en-pl	0.7793	0.7442	0.738	0.7426	0.7323	0.7665	0.7491	0.7318	0.731	0.7303	0.7337	0.7365	0.7429
en-uk	0.7711	0.7107	0.7193	0.7248	0.7144	0.7516	0.7199	0.7157	0.7145	0.7135	0.718	0.7216	0.7246
pl-de	0.7489	0.665	0.6772	0.6861	0.6652	0.6556	0.6808	0.6841	0.6853	0.6861	0.6914	0.6961	0.6852
pl-en	0.7562	0.7303	0.7343	0.7365	0.7319	0.7326	0.7345	0.73	0.7314	0.7325	0.7344	0.7361	0.7351
pl-uk	0.7476	0.6776	0.6651	0.6542	0.6775	0.7031	0.6914	0.6584	0.6647	0.6696	0.6773	0.6838	0.6809
uk-de	0.7242	0.6277	0.6496	0.6603	0.6347	0.6404	0.6488	0.6595	0.6623	0.6644	0.6699	0.6745	0.6597
uk-en	0.7595	0.7129	0.724	0.7281	0.7185	0.7277	0.7221	0.7231	0.7257	0.7278	0.7305	0.7329	0.7277
uk-pl	0.7245	0.6629	0.6458	0.6416	0.6546	0.6845	0.6746	0.6455	0.6511	0.6555	0.6622	0.6679	0.6642
alphanumeric	0.7507	0.699	0.7037	0.707	0.7005	0.7079	0.7089	0.7085	0.7093	0.71	0.7134	0.7164	0.7113
email	0.7338	0.6719	0.6821	0.6869	0.6761	0.6925	0.6847	0.6883	0.688	0.6878	0.6918	0.6952	0.6899
emoji	0.7606	0.7045	0.7041	0.7079	0.6989	0.7207	0.7121	0.7069	0.709	0.7107	0.715	0.7186	0.7141
iban	0.7616	0.7043	0.7089	0.709	0.7072	0.7194	0.7154	0.7113	0.7125	0.7134	0.7176	0.7211	0.7168
ip	0.7777	0.7278	0.7321	0.735	0.7278	0.7382	0.7378	0.7357	0.7368	0.7375	0.741	0.7439	0.7393
isbn	0.772	0.7178	0.7238	0.7276	0.7205	0.7291	0.7292	0.7286	0.7295	0.7301	0.7338	0.7369	0.7316
phone	0.766	0.7127	0.7147	0.7175	0.7119	0.7236	0.7228	0.7183	0.7193	0.72	0.7237	0.7269	0.7231
social	0.769	0.7082	0.7171	0.722	0.7113	0.7285	0.7194	0.7215	0.7229	0.7239	0.7279	0.7313	0.7253
url	0.7471	0.6849	0.6963	0.7005	0.6903	0.701	0.6973	0.7014	0.7025	0.7034	0.7074	0.7107	0.7036
macro avg. by model	0.7598	0.7035	0.7092	0.7126	0.7049	0.7179	0.7142	0.7103	0.7117	0.7127	0.7168	0.7202	

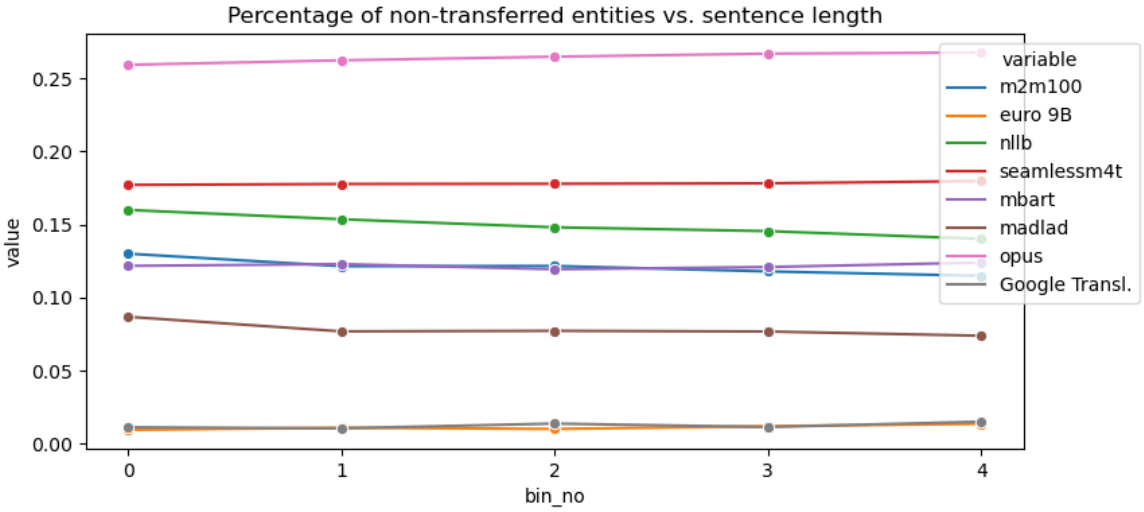


Figure 5: Percentage of entities not matched in target sentences among different lengths.

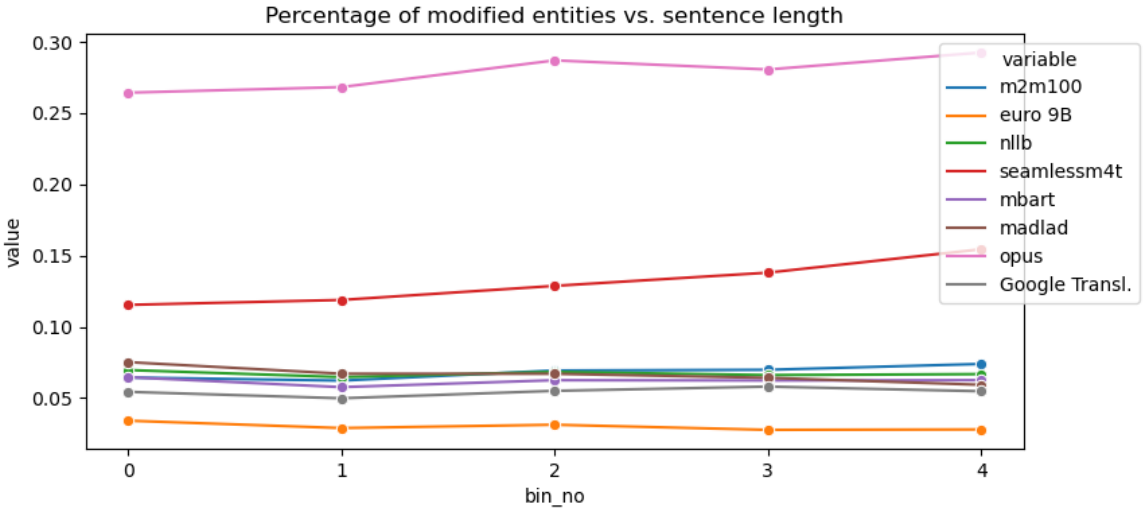


Figure 6: Percentage of modified entities among different lengths.

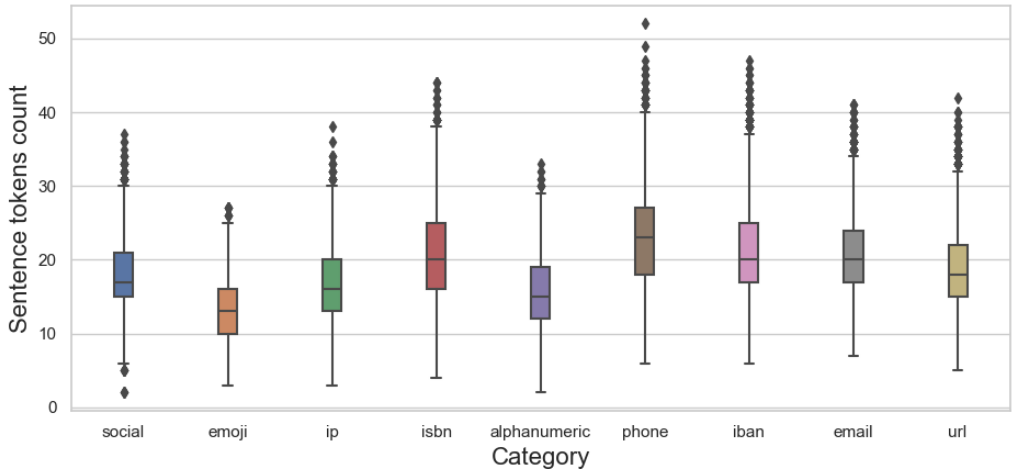


Figure 7: Number of NLTK (Bird and Loper, 2004) tokens per category. Sentences expressed in all languages were collected together for this analysis.

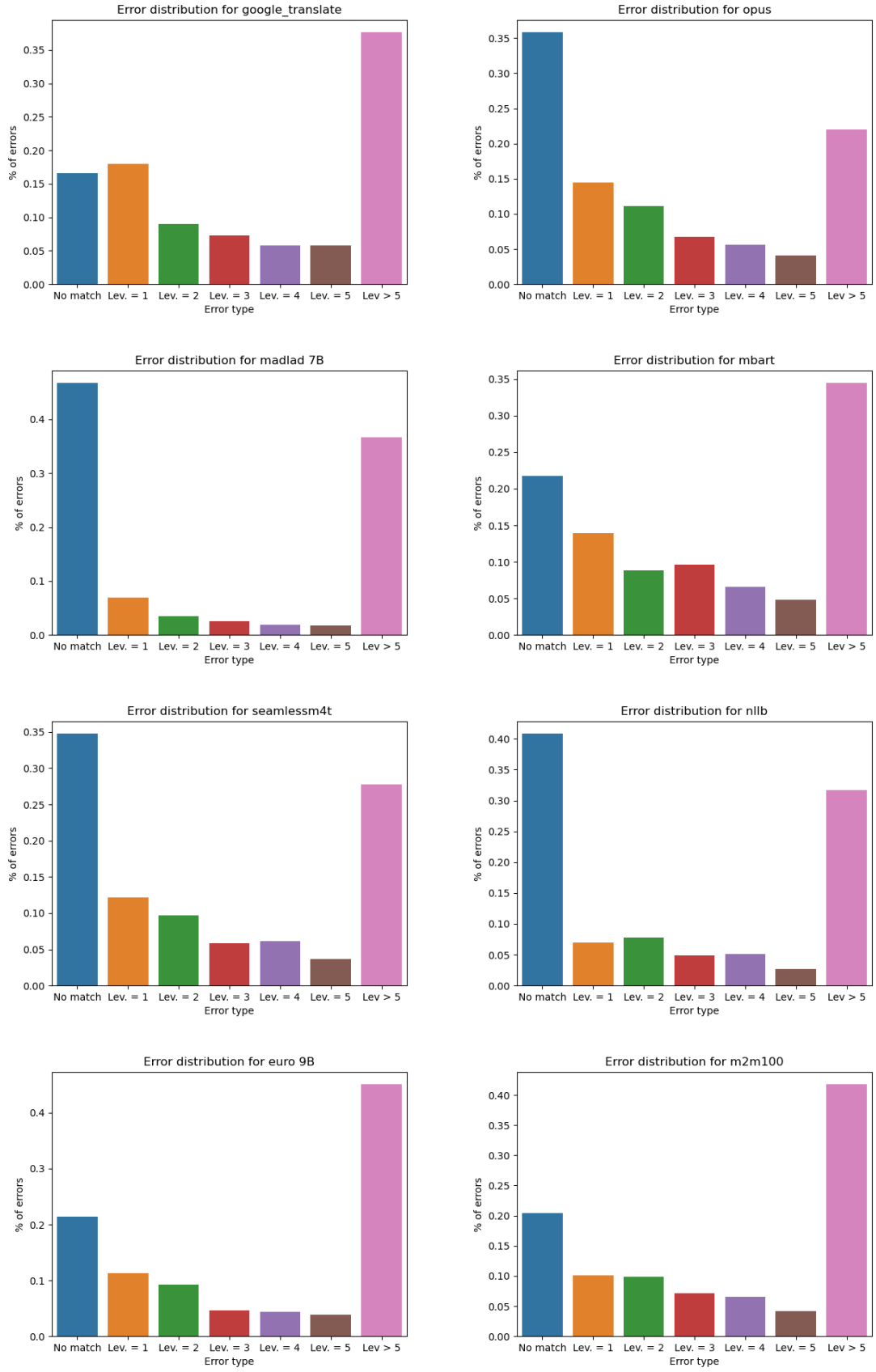


Figure 8: Errors distributions without emoji category

Table 10: Examples of errors made by Google Translate (verified on 01.01.2025)

comment	**IBAN modified**, pl → de
srcEntity	**PL60109010000000000000000000**
tgtEntity	**PL60109010000000000000000**
srcText	Panie profesorze Janie Kowalski, może przesłać mi nową listę prac do sprawdzenia na konto o numerze PL60109010000000000000000000?
tgtText	Professor Jan Kowalski, können Sie mir eine neue Liste der zu prüfenden Werke an die Kontonummer PL60109010000000000000000 senden?
comment	**Translated social handler**, en → de
srcEntity	**@klimatyzacja**
tgtEntity	**@airconditioning**
srcText	Przechodząc przez park, nagle usłyszałam @klimatyzacja śpiewającą piosenkę o letnim słońcu.
tgtText	Als ich durch den Park spazierte, hörte ich plötzlich @airconditioning ein Lied über die Sommersonne singen.
comment	**Changed characters in alphanumeric**, en → de
srcEntity	**tenotypic123CBSprk**
tgtEntity	**tenotypisch123CBSprk**
srcText	Gniazdo pająka, o symbolu tenotypic123CBSprk, wisiało pod niebem usianym szumami.
tgtText	Das Spinnennest, Symbol tenotypisch123CBSprk, hing unter einem mit Lärm übersäten Himmel.
comment	**Changed characters in URL**, de → en
srcEntity	**www.irgendwohin.com**
tgtEntity	**www.somewhere.com**
srcText	Die Katze las www.irgendwohin.com vor dem Frühstück und kraulte verschmitzt um Aufmerksamkeit.
tgtText	The cat read www.somewhere.com before breakfast and playfully scratched for attention.
comment	**Modified e-mail** de → en
srcEntity	**liebevollchenpinguin@aya.at**
tgtEntity	**lovingchenpinguin@aya.at**
srcText	Die sprechende Mandarine geschickt ein Bild an liebevollchenpinguin@aya.at.
tgtText	The talking mandarin sent a picture to lovingchenpinguin@aya.at.
comment	**IP removed**, pl → de
srcEntity	**192.168.1.108**
tgtEntity	**NO ENTITY DETECTED**
srcText	Wiatr gonił pożółkłe liście, aż mu się zadało z 192.168.1.108 i spróbowało wziąć pod nie chwytem.
tgtText	Der Wind verfolgte die vergilbten Blätter, bis er müde wurde und versuchte, sie zu ergreifen.
comment	**Dropped Phone number**, de → pl
srcEntity	**49 030 1234567890**
tgtEntity	**NO ENTITY DETECTED**
srcText	Die alten Ratten spielten Karten und diskutierten leidenschaftlich laut vor Telefonnummer +49 030 1234567890 verband 123politisch.
tgtText	Stare szczury grały w karty i głośno i namiętnie dyskutowały o polityce.
comment	**ISBN dropped** pl → de
srcEntity	**978-83-12-34567-8**
tgtEntity	**NO ENTITY DETECTED**
srcText	Podczas lekcji astronomii, Paweł natknął się na książkę o istocie czasoprzestrzeni, której ISBN 978-83-12-34567-8 zdradził tajemnicę kosmicznej harmonii.
tgtText	Während einer Astronomiestunde stieß Paweł auf ein Buch über die Natur der Raumzeit, das das Geheimnis der kosmischen Harmonie enthüllte.

Intrinsic vs. Extrinsic Evaluation of Czech Sentence Embeddings: Semantic Relevance Doesn't Help with MT Evaluation

Petra Barančíková and **Ondřej Bojar**
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{barancikova,bojar}@ufal.mff.cuni.cz

Abstract

In this paper, we compare Czech-specific and multilingual sentence embedding models through intrinsic and extrinsic evaluation paradigms. For intrinsic evaluation, we employ Costra, a complex sentence transformation dataset, and several Semantic Textual Similarity (STS) benchmarks to assess the ability of the embeddings to capture linguistic phenomena such as semantic similarity, temporal aspects, and stylistic variations. In the extrinsic evaluation, we fine-tune each embedding model using COMET-based metrics for machine translation evaluation.

Our experiments reveal an interesting disconnect: models that excel in intrinsic semantic similarity tests do not consistently yield superior performance on downstream translation evaluation tasks. Conversely, models with seemingly over-smoothed embedding spaces can, through fine-tuning, achieve excellent results. These findings highlight the complex relationship between semantic property probes and downstream task, emphasizing the need for more research into "operationalizable semantics" in sentence embeddings, or more in-depth downstream tasks datasets (here translation evaluation).

1 Introduction

Machine translation (MT) evaluation has advanced significantly in recent years, finally moving beyond traditional surface-level metrics like BLEU (Papineni et al., 2002) towards more sophisticated approaches based on neural networks and contextualized embeddings.

State-of-the-art MT evaluation metrics such as COMET (Rei et al., 2022b) and BLEURT (Sellam et al., 2020) use sentence embeddings to better capture semantic similarity between translations and references, achieving much higher correlation with human judgments than traditional metrics.

However, the rapid development of new embedding models presents MT researchers with a challenging choice. Although multilingual models such as LaBSE (Feng et al., 2022) and XLM-RoBERTa (Conneau et al., 2019) have shown strong cross-lingual capabilities, there are also language-specific models that claim superior performance for a selected language. For morphologically rich languages like Czech, it remains unclear whether these specialized sentence embeddings offer advantages over multilingual alternatives when used in MT evaluation.

In this paper, we examine the evaluation of English-to-Czech machine translation and compare several state-of-the-art Czech-specific models against multilingual models using both intrinsic evaluation and extrinsic evaluation. To this end, we see the task of machine translation evaluation (MTE) and quality estimation (QE), i.e. MTE without professionally translated reference sentences, as methods for extrinsic evaluation of sentence embeddings. For intrinsic evaluation, we assess how well the examined sentence embeddings reflect semantic properties exemplified in two datasets: Costra (Barančíková and Bojar, 2020) and Semantic Textual Similarity (STS, Bednář et al., 2024). In sum, our goal is to understand whether the performance of a model in intrinsic semantic tasks correlates with its usability for MT evaluation, potentially simplifying the selection of embeddings.

2 Related Work

Several studies have raised concerns about the use of STS as an evaluation metric. For instance, Reimers et al. (2016), Eger et al. (2019), and Zhelezniak et al. (2019) argue that, while STS can capture certain semantic similarities, it does not reliably predict how effective sentence representa-

Proceedings of Machine Translation Summit XX Volume 1, pages 265–275, June 23-27, 2025

tions will be for downstream tasks. These works highlight how STS tasks often encourage surface-level heuristics or oversimplified semantic similarity patterns that may not generalize to more complex applications like entailment or paraphrasing detection.

To address these limitations, new intrinsic evaluation methods such as EvalRank (Wang et al., 2022) and SentBench (Xiaoming et al., 2023) have been proposed, both of which exhibit a stronger correlation with extrinsic evaluation measures. These benchmarks evaluate sentence representations through information retrieval, sentence ordering, and probing tasks, offering a more holistic view of embedding quality that aligns better with actual downstream task performance.

It is important to note that these previous experiments did not specifically focus on machine translation evaluation, which seems to be very close to STS—it also involves comparing pairs of sentences to assess their semantic closeness. Cífka and Bojar (2018) report a negative correlation between the translation quality of Transformer models measured by BLEU and the semantic properties (assessed using STS) of the sentence embeddings derived from the Transformer model. In contrast, Libovický and Madhyastha (2019) demonstrate a strong positive correlation between STS performance and translation quality for both Transformer and RNN-based models.

More recently, Freitag et al. (2022) have advocated for the use of semantic-aware metrics such as BERTScore (Zhang et al., 2020) and COMET (Rei et al., 2020) in MT evaluation, showing that these outperform BLEU in correlating with human judgments. These models incorporate contextual embeddings and often exhibit closer alignment with human-perceived meaning, bringing MT evaluation closer to the goals of intrinsic semantic understanding.

3 Models

For our experiments, we used several state-of-the-art sentence embedding models, employing both Czech-specific and multilingual variants. The Czech encoders include three *base*-size Transformer architectures, each using masked language modeling as their primary pretraining objective—CZERT-b-cased (`Czert`, Sido et al., 2021), FERNET-C5 (`FERNET`, Lehečka and Švec, 2021) and `RobeCzech` (Straka et al., 2021).

To provide a broader comparative analysis, we also experimented with multilingual sentence embedding models trained on datasets that contain Czech texts. These include `LaBSE` (Feng et al., 2022), a model generating language-agnostic representations for more than one hundred languages with remarkable cross-lingual alignment, since its training objective was machine translation. Furthermore, we evaluated two *large* models: XLM-RoBERTa-large (`XLM-R`, Conneau et al., 2019) and multilingual-e5-large (`mE5`, Wang et al., 2024), a model pretrained using a contrastive learning approach on a diverse range of tasks, including natural language inference and question answering across multiple languages.

As a baseline model, we employed a BERT architecture (Devlin et al., 2019) with randomly initialized weights. The only component inherited from the pretrained *'bert-base-multilingual-cased'* model is the tokenizer. This means that while the model processes input according to the tokenization patterns learned from multilingual data, it does not benefit from any pretrained language representations. We refer to this configuration as `random BERT`. This setup isolates and assess the contribution of the tokenizer alone, establishing a lower performance bound and offering a meaningful point of comparison to evaluate the benefits of pretraining.

4 Intrinsic Evaluation

We first evaluate the embeddings using a series of semantic benchmarks to determine their ability to accurately capture various semantic properties of a sentence.

4.1 Costra

As the first dataset for intrinsic evaluation, we used the Costra[1] dataset (Barančíková and Bojar, 2020). It was created manually, specifically to test the quality of Czech embeddings, focusing on complex transformations of sentences beyond standard paraphrasing or simple word-level changes. The sentence embeddings are tested across the following six categories:

- **Basic**: evaluates whether paraphrases are positioned closer together in embedding space compared to transformations that significantly alter the meaning of the original sentence.

[1] https://github.com/barancik/costra

- **Modality**: measures whether paraphrases are more similar to their original sentence than transformations that change the sentence's modality (e.g., possibility or prohibition).

- **Time, Style, Generalization, Opposite**: these categories test embeddings' ability to reflect linear ordering of sentence variations (e.g., from the least general to the most general) as proposed by annotators.

Each category is scored on a scale from 0 (worst) to 1 (best), reflecting the proportion of Costra sentence triplets for which the relations in the sentence vector space align with human annotations. For example, consider a triplet consisting of a seed sentence S, its paraphrase P, and its opposite sentence O. Ideally, the cosine similarity S_C should satisfy $S_C(S,P) > S_C(S,O)$ and $S_C(S,P) > S_C(P,O)$, indicating that the model correctly identifies the paraphrase closer to the seed sentence than the opposite sentence.

The results are presented in Table 1, with the overall Costra score calculated as the arithmetic mean across all six categories. In particular, the evaluation shows that all sophisticated models failed to outperform randomly generated embeddings[2] in the first two categories, **Basic** and **Modality**. In fact, these categories were designed to be particularly challenging, including comparisons of paraphrases with substantial lexical variation and sentences that, despite the close lexical similarity to a paraphrase, differ significantly in meaning. These results suggest that all models were fooled by surface-level similarity, making randomly generated embeddings the overall winner in these two categories. Consequently, it is impossible to distinguish whether slight improvements in these categories can be attributed to model quality or to randomness.

To address this limitation, we introduce the **Costra–** score, calculated as the average of the four remaining categories: **Time**, **Style**, **Generalization**, and **Opposite**. However, the **Costra–** scores revealed only marginal differences across models. The smallest model, `SimCSE`, slightly outperformed its counterparts but the improvement was not substantial. In fact, the models performed only marginally better than the `random BERT` model, suggesting limited success in capturing phenomena tested in the Costra dataset, such as linearity of time or generalization. Several models, including large `XLM-R`, even performed worse than `random BERT`.

4.2 Semantic Textual Similarity

Table 2 presents the results of our evaluation of sentence embeddings on the Semantic Textual Similarity (STS) task. Performance is measured using the automated evaluation tool[3] provided by Bednář et al. (2024). This tool computes similarity for pairs of sentences in three STS datasets. For precomputed sentence embeddings, it explores different embedding similarity metrics including cosine similarity, dot product, and Manhattan distance. Additionally, it applies various sentence embeddings pooling strategies and selects the highest average score as the final result.

Interestingly, the results are consistent with findings from Costra, with `SimCSE` being the overall best performing model, followed by `mE5` and `LaBSE` in the next two positions. Surprisingly, `XLM-R`, despite being a powerful multilingual model, may not be well-optimized for Czech-specific STS tasks, ranking last in the evaluation, performing even worse than `random BERT`.

5 Extrinsic Evaluation—MTE and QE

Extrinsic evaluation utilizes sentence embeddings as feature vectors for machine learning algorithms in downstream NLP tasks—MTE and QE in our tasks. It serves well to choose the best method for a particular task but not as an absolute metric of embedding quality, as the performance of the embeddings does not correlate across different tasks (Bakarov, 2018).

5.1 Data

In the following experiments, we utilize datasets from the Workshop on Machine Translation (WMT), selecting data from English-to-Czech translations. These datasets include English source sentences, Czech hypotheses (i.e., machine translated outputs), Czech reference sentences, and the human translation quality scores collected using the Direct Assessment (DA) method (Graham et al., 2013) and subsequently z-normalized.

Data from WMT17 to WMT19 (Bojar et al., 2017, 2018; Barrault et al., 2019) were used to

[2]Not to be confused with `random BERT`, we evaluated Costra also using completely *random vectors*.

[3]https://github.com/seznam/czech-semantic-embedding-models

				Costra–					
		Costra							
Embeddings	**Size**	**Basic**	**Mod.**	**Time**	**Style**	**Gen.**	**Opp.**	**Costra**	**Costra–**
SimCSE	256	0.20	0.35	**0.74**	0.63	0.73	**0.78**	**0.57**	**0.72**
mE5	1,024	0.24	0.34	0.71	0.62	**0.75**	0.77	**0.57**	0.71
LaBSE	768	0.20	0.26	0.71	0.63	**0.75**	0.75	0.55	0.71
RetroMAE	256	0.06	0.06	0.69	0.63	0.70	0.76	0.48	0.70
RobeCzech	768	0.15	0.13	0.69	**0.65**	0.69	0.75	0.51	0.70
random BERT	768	0.08	0.06	0.65	0.60	0.72	0.73	0.47	0.68
Czert	768	0.31	0.35	0.66	0.64	0.69	0.69	0.56	0.67
XLM-R	1,024	0.16	0.11	0.65	0.61	0.67	0.68	0.48	0.65
FERNET	768	0.33	0.38	0.65	0.61	0.63	0.68	0.54	0.64
random vectors	256	**0.50**	**0.51**	0.49	0.50	0.49	0.50	0.50	0.50

Table 1: This Table presents the results of intrinsic evaluation using the Costra dataset. The Costra score ranges from 0 (worst) to 1 (best) in each category. The overall Costra score is calculated as the arithmetic mean across all categories. Costra– represents the mean score excluding the first two categories (Basic and Mod.), as these categories appear excessively challenging for all pretrained encoders evaluated.

Embeddings	**avg. similarity**
SimCSE	**87.83**
LaBSE	82.91
mE5	78.39
RetroMAE	76.30
Czert	74.79
RobeCzech	70.28
FERNET	65.46
random BERT	60.48
XLM-R	57.88

Table 2: Results of intrinsic evaluation on three STS datasets.

train the COMET estimators (Rei et al., 2020). The validation of the models was performed on the WMT20 dataset (Barrault et al., 2020), and the performance of the models was tested using the WTM21 (Akhbardeh et al., 2021) and WMT22 (Kocmi et al., 2022) datasets.

5.2 MTE Baseline Approach

Before fine-tuning the sentence embedding models for machine translation evaluation, we conducted a preliminary analysis to assess their default ability to evaluate translation quality. Specifically, we examined Pearson's correlation between human judgments and the cosine similarities computed between (i) a hypothesis and a reference translation and (ii) a hypothesis and a source sentence. We expected high cosine similarity for multilingual models, reflecting their ability to capture cross-lingual semantic relationships, whereas Czech-specific models—lacking such cross-lingual information—were anticipated to have random similarity scores.

Furthermore, we examined the intrinsic quality of the embedding spaces by measuring the cosine similarity between the source and reference embeddings. We also performed a random shuffling experiment designed to evaluate the discriminative ability of the embeddings.

The results presented in Table 3 reveal that even without fine-tuning, a slight correlation between human judgments and cosine similarity of hypotheses and references is observable in certain models—particularly mE5, RetroMAE, and SimCSE. However, contrary to expectations, this does not hold for source sentences; no relationship was detected between human evaluation scores and the cosine similarity computed between a translated sentence and its source sentence, even among the multilingual models.

The analysis of the embedding space via similarity between source and reference sentences provides further insights. In line with our hypothesis, XLM-R exhibits a near perfect similarity between the source and reference sentences, indicative of a tightly clustered or language-agnostic representation; however, the same holds for random BERT.

To further investigate this behavior, we repeated the experiment using random shuffle of source and

Sentence Embeddings	WMT21 test set			WMT22 test set			
	$\rho_{H,R}$	$\rho_{H,S}$	$S_C(S,R)$	$\rho_{H,R}$	$\rho_{H,S}$	$S_C(S,R)$	$S_C(S_R,R_R)$
mE5	**0.29**	0.04	0.89	0.26	0.01	0.90	0.75
RetroMAE	0.26	-0.10	0.76	**0.27**	**0.09**	0.76	0.69
SimCSE	0.24	**0.13**	0.85	0.25	0.05	0.82	0.09
Czert	0.20	-0.03	0.63	0.18	-0.06	0.62	0.52
XLM-R	0.17	-0.08	1.00	0.05	-0.10	1.00	0.99
RobeCzech	0.15	-0.16	0.92	0.11	-0.06	0.91	0.89
LaBSE	0.11	0.03	0.89	0.19	0.06	0.88	0.31
FERNET	0.07	-0.11	0.45	0.11	-0.03	0.40	0.35
random BERT	0.06	-0.16	0.99	0.03	-0.20	0.98	0.98

Table 3: Results for baseline MTE approach—using sentence embeddings for direct evaluation without fine-tuning. $\rho_{H,R}$ represents Pearson correlation between human quality assessments and the cosine similarity between the translation hypothesis and the reference translation, while $\rho_{H,S}$ shows the correlation between human judgments and the cosine similarity between the hypothesis and the source sentence. $S_C(S,R)$ represents cosine similarity between references and sources. The last column represents cosine similarity between randomly shuffled source and reference sentences averaged over 100 runs.

reference sentences; see the last column of Table 3. The similarity remained perfect for both XLM-R and random BERT even on shuffled pairs, indicating an overly invariant embedding space, where even pairs of semantically unrelated sentences tend to cluster together. This *over-smoothing* reduces the model's capacity to distinguish subtle differences that are essential for evaluating translation quality. In such cases, even bad translations can receive high similarity scores, lowering the correlation with human judgment. This also explains the poor performance of XLM-R in our intrinsic evaluation task, especially in STS (Table 2). More broadly speaking, it casts doubts on any results based on the direct similarity of XLM-R embedding vectors in the Czech language, given that XLM-R assigns similar vectors to random Czech sentences.

5.3 Models fine-tuning for MTE and QE

For all sentence encoders, we fine-tuned two COMET-based estimators (CE, Rei et al., 2020), one for machine translation evaluation using reference sentences and the other for quality estimation without reference sentences. The COMET models use a dual-encoder architecture: the source sentence, reference translation, and hypothesis are each processed independently using transformer encoder models followed by two hidden layers of sizes 3072 (resp. 2048 for QE) and 1024.

We used the default training settings with the AdamW optimizer ($1.5 \cdot 10^{-5}$ for the regression layers and $1.0 \cdot 10^{-6}$ for the encoder) and a layer-wise decay of 0.95. To preserve encoder generalization, the embeddings were frozen for the first 0.3 epochs. Both models used mixed-layer pooling with a sparsemax-based transformation before pooling and were optimized with mean squared error loss (using a dropout of 0.1). Training was conducted over five epochs, and we selected the checkpoint with the highest Kendall's tau validation on a held-out validation dataset.

These settings were applied consistently across all models without extensive hyperparameter tuning. In total, we trained a total of 18 COMET estimators. To avoid confusion with the original embeddings, we refer to a trained COMET estimator for given embeddings X as to $\mathrm{CE}_{\mathrm{MTE}}$(X) for machine translation with reference sentences and $\mathrm{CE}_{\mathrm{QE}}$(X) for the referenceless quality estimation metric (e.g., for the Czert embeddings, we use $\mathrm{CE}_{\mathrm{MTE}}$(Czert) and $\mathrm{CE}_{\mathrm{QE}}$(Czert), respectively).

5.4 Results of MTE and QE evaluation

We compare the performance of trained evaluation metrics at the system level with traditional string-matching MTE metrics. Specifically, we include BLEU (Papineni et al., 2002), TER (Snover et al., 2006), and ChrF (Popović, 2015), using their default configurations as implemented in SacreBLEU (Post, 2018). Additionally, we employ METEOR-NEXT (Denkowski and Lavie, 2010), a metric that include paraphrase support, on both system and segment levels.

Furthermore, we compute scores using the official pretrained COMET models for machine translation evaluation, namely wmt22-comet-da (Rei

	system-level		segment-level	
MTE metrics	2021	2022	2021	2022
CE_{MTE}(FERNET)	**0.98**	**0.97**	0.60	0.47
wmt22-comet-da	0.97	0.93	**0.66**	**0.51**
CE_{MTE}(Czert)	0.96	0.93	0.57	0.43
CE_{MTE}(XLM-R)	0.96	0.93	0.62	0.47
CE_{MTE}(RobeCzech)	0.97	0.92	0.58	0.44
CE_{MTE}(mE5)	0.96	0.92	0.59	0.46
METEOR-NEXT	**0.98**	0.84	0.24	0.21
chrF2	0.97	0.84	-	-
CE_{MTE}(RetroMAE)	0.91	0.82	0.43	0.34
CE_{MTE}(LaBSE)	0.89	0.79	0.56	0.45
CE_{MTE}(SimCSE)	0.96	0.74	0.44	0.37
1-TER	0.95	0.60	-	-
BLEU	0.94	0.54	-	-
CE_{MTE}(random BERT)	0.35	-0.35	0.23	0.22

	system-level		segment-level	
QE metrics	2021	2022	2021	2022
CE_{QE}(FERNET)	**0.98**	**0.96**	0.60	0.46
wmt22-cometkiwi-da	0.95	0.91	**0.67**	**0.49**
CE_{QE}(XLM-R)	0.96	0.88	0.63	**0.49**
CE_{QE}(RobeCzech)	0.97	0.87	0.58	0.39
CE_{QE}(Czert)	0.95	0.86	0.57	0.39
CE_{QE}(mE5)	0.93	0.76	0.59	0.45
CE_{QE}(LaBSE)	0.83	0.39	0.54	0.40
CE_{QE}(RetroMAE)	0.64	0.15	0.39	0.23
CE_{QE}(random BERT)	0.47	-0.19	0.26	0.20
CE_{QE}(SimCSE)	0.12	-0.92	0.38	0.24

Table 4: Correlations between human scores and evaluation metrics, including both fine-tuned COMET-based metrics and traditional metrics, computed at the system and segment levels.

et al., 2022a), and for quality estimation, specifically wmt22-cometkiwi-da (Rei et al., 2022b). These COMET models extend beyond a simple trained COMET estimator, as they incorporate an ensemble approach combining a COMET estimator trained on DA data and sequence predictors trained on MQM annotations.

The results in Table 4 indicate a clear advantage for COMET-based evaluation metrics over traditional metrics in MTE. In the system-level analysis, the COMET variants CE_{MTE}(FERNET) and CE_{QE}(FERNET) achieved consistently remarkably high correlation outperforming even the official COMET ensemble metrics – wmt22-comet-da and wmt22-cometkiwi-da, which were the top-performing metrics at the segment level.

In contrast, classical metrics, although competitive in 2021, showed significant performance degradation in 2022. CE_{MTE}(random BERT) failed completely, highlighting the importance of using pretrained sentence embeddings, even though CE_{QE}(random BERT) outperformed CE_{QE}(SimCSE), even though SimCSE was the best performing encoder in intrinsic evaluation.

Another interesting observation is the small difference in correlations of the top performing embeddings between MTE and QE. The correlation of CE_{MTE}(FERNET) and CE_{QE}(FERNET) is practically equal at both the system and segment levels, as if these metrics no longer have use for reference translations. This is consistent with recent research showing that reference-free evaluation has become competitive with reference-based evaluation (Rei et al., 2021) or even outperforms it (Moosa et al., 2024).

6 Results and Discussion

When comparing the results of MTE and QE with those of the intrinsic evaluation tasks, we can observe an interesting inversion. Although both evaluation approaches aim to capture semantic similarity, the performance of the embeddings changed significantly after fine-tunings. Specifically, `XLM-R` and `FERNET` embeddings, which performed poorly in intrinsic evaluation, became the best performing MTE and QE metrics. In contrast, `SimCSE`, which dominated intrinsic evaluations, ranks among the worst performing metrics in MTE and QE. These results are in line with related research (Section 2), which shows that STS performance may not accurately predict effectiveness in downstream tasks.

There are several plausible hypotheses that might explain these discrepancies. Let us at least mention them here— unfortunately, their thorough testing is beyond the scope of this article.

First, `XLM-R` and `FERNET` might perform poorly in intrinsic tasks because their representation spaces are not tuned for fine-grained semantic differences. However, when fine-tuned on a translation quality task, the model might learn to emphasize those aspects of the embedding space that are important for distinguishing translation quality.

The fine-tuning process for COMET-based evaluation might be effectively reconfiguring the `XLM-R` embedding space, transforming its initially oversmoothed representations into task-specific features that are highly discriminative for translation quality. Although `XLM-R` raw embeddings appear to be all clustered together (see Table 3), the fine-tuning may introduce transformations that reweight and separate the dimensions relevant for capturing translation errors. In contrast, `SimCSE` embeddings, which are already optimized for intrinsic semantic discrimination, might leave less room for adjustments necessary to learn the new training objective.

We should also not forget about the different embedding sizes, which played an important role in the observed behavior. The *small* embeddings—`SimCSE` and `RetroMAE`—were among the worst performing COMET estimators. Large embeddings, such as those produced by `XLM-R`, offer a higher-dimensional space that can capture more nuanced semantic and syntactic features. When fine-tuning with the COMET estimator—which adds two hidden layers with sizes 3072 (resp. 2048) and 1024—the richer representation provided by larger embeddings could allow the model to extract and emphasize the translation-specific signals more effectively.

Interestingly, we can see not too much difference between the monolingual vs. multilingual embedding performances—they seem to be equally represented among the best performing embeddings in both intrinsic and extrinsic tasks. The size of the embeddings seem not to matter in the intrinsic tasks—the top 3 best performing embeddings (`SimCSE`, `LaBSE` and `mE5`) are *small*, *base* and *large*, respectively.

The correlation analysis between different evaluation methodologies, visualized by heatmaps in Figure 1, reveals interesting patterns of how different evaluation methodologies relate to each other. These patterns provide valuable insight into the reliability and consistency of various embedding evaluation approaches.

The heatmaps highlight a strong alignment among all intrinsic tasks (Costra–, STS, and MTE baselines). Moreover, there is a strong correlation between the segment-level and the system-level metrics, indicating that aggregated segment scores provide reliable system-level insights. In particular, we observe strong correlations between segment-level metrics (*segment MTE* and *segment QE* showing correlations of 0.97 and 0.91 for 2021 and 2022 respectively), suggesting that these evaluation approaches capture similar aspects of translation quality despite their methodological differences.

However, one of the most striking findings is the weak and sometimes even negative correlation between intrinsic evaluation metrics (Costra–, STS) and the system-level quality estimation scores *system QE*. This discrepancy is particularly evident in the 2022 data, where Costra– shows a negative correlation (-0.52) with *system QE*, challenging the assumption that better semantic representation capabilities necessarily translate to improved MT evaluation performance.

These findings indicate that intrinsic measures, while useful for general semantic similarity, may not sufficiently reflect translation-specific nuances required for MTE or QE. Consequently, intrinsic criteria alone appear inadequate for selecting optimal sentence embeddings for these specific tasks. Further research is needed to identify intrinsic evaluation methods that better capture the subtleties relevant to machine translation. Additionally, it would be valuable to explore in more detail the types of errors penalized in manual MT quality assessments

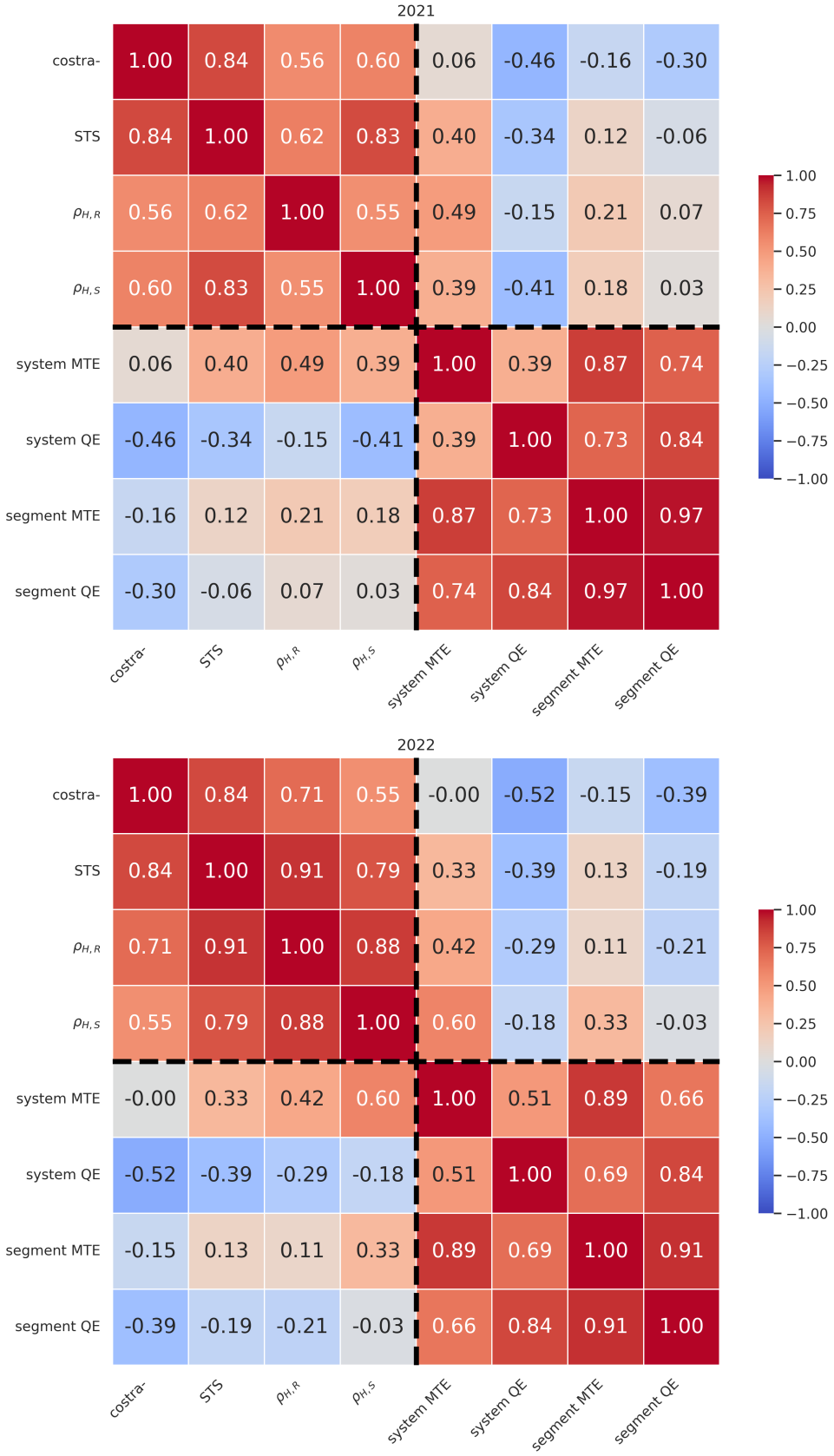


Figure 1: Correlation heatmaps for different method of embedding evaluations. The heatmaps are intentionally kept in a rectangular shape to emphasize the mismatch in correlation patterns between intrinsic evaluation (Costra–, STS, $\rho_{H,R}$, $\rho_{H,S}$) and extrinsic evaluation (system MTE/QE and segment MTE/QE).

to determine whether these errors predominantly concern sentence meaning or other aspects that should be preserved in translation.

7 Conclusion and Future Work

We experimented with several evaluation methods for both Czech and multilingual sentence embeddings, considering intrinsic semantic tasks and downstream application in machine translation evaluation and quality estimation. Our key findings include the following:

- **Intrinsic vs. Extrinsic Discrepancy:** The lack of correspondence between the intrinsic and extrinsic metrics used in our experiments suggests that intrinsic evaluation methods employing these metrics cannot reliably predict a model's performance in MT evaluation tasks. This finding suggests the need for better targeted intrinsic evaluation approaches that reflect downstream application requirements (Figure 1).
- **Temporal Stability:** The stability of the correlations over time between the segment-level metrics provides encouraging evidence for the reliability of these evaluation approaches.
- **Language-Specific vs. Multilingual Models:** There are no strong differences in performance between language-specific and multilingual models. Both categories are comparably represented among the top-performing models in intrinsic and extrinsic tasks.
- **Model Size Might Matter:** In contrast to intrinsic tasks, fine-tuning embeddings for MTE/QE reveals that model size does matter, with the *small* embeddings consistently showing poor performance.

In future work, we intend to replicate these experiments across multiple languages to investigate whether the observed behavior is specific to the Czech language or if it generalizes to other languages. In addition, we plan to conduct a more thorough analysis to better understand the underlying reasons for the differences in performance between the evaluation methods.

Acknowledgments

The authors acknowledge the support of the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO and 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." This work has been conducted using data 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, Online. Association for Computational Linguistics.

Amir Bakarov. 2018. A Survey of Word Embeddings Evaluation Methods. *CoRR*, abs/1801.09536.

Petra Barančíková and Ondřej Bojar. 2020. COSTRA 1.0: A Dataset of Complex Sentence Transformations. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3535–3541, Marseille, France. European Language Resources Association.

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.

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.

Jiří Bednář, Jakub Náplava, Petra Barančíková, and Ondřej Lisickỳ. 2024. Some Like It Small: Czech Semantic Embedding Models for Industry Applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22734–22742.

Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, and Julia Kreutzer, editors. 2017. *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*. 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, Belgium, Brussels. Association for Computational Linguistics.

Ondřej Cífka and Ondřej Bojar. 2018. Are BLEU and Meaning Representation in Opposition? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1362–1371, Melbourne, Australia. 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. *CoRR*, abs/1911.02116.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support for Five Target Languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 339–342, Uppsala, Sweden. 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.

Steffen Eger, Andreas Rücklé, and Iryna Gurevych. 2019. Pitfalls in the Evaluation of Sentence Embeddings. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 55–60, Florence, Italy. 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.

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.

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.

Jan Lehečka and Jan Švec. 2021. Comparison of Czech Transformers on Text Classification Tasks. In *Statistical Language and Speech Processing*, pages 27–37, Cham. Springer International Publishing.

Jindrich Libovický and Pranava Madhyastha. 2019. Probing Representations Learned by Multimodal Recurrent and Transformer Models. *CoRR*, abs/1908.11125.

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*.

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.

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, 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. 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.

Nils Reimers, Philip Beyer, and Iryna Gurevych. 2016. Task-Oriented Intrinsic Evaluation of Semantic Textual Similarity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 87–96, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Jakub Sido, Ondřej Pražák, Pavel Přibáň, Jan Pašek, Michal Seják, and Miloslav Konopík. 2021. Czert – Czech BERT-like Model for Language Representation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1326–1338, Held Online. INCOMA Ltd.

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.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. RobeCzech: Czech RoBERTa, a Monolingual Contextualized Language Representation Model. In *Text, Speech, and Dialogue*, pages 197–209, Cham. Springer International Publishing.

Bin Wang, C.-C. Jay Kuo, and Haizhou Li. 2022. Just Rank: Rethinking Evaluation with Word and Sentence Similarities. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6060–6077, Dublin, Ireland. Association for Computational Linguistics.

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*.

Liu Xiaoming, Lin Hongyu, Han Xianpei, and Sun Le. 2023. SentBench: Comprehensive Evaluation of Self-Supervised Sentence Representation with Benchmark Construction. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 813–823, Harbin, China. Chinese Information Processing Society of China.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Hammerla. 2019. Correlation Coefficients and Semantic Textual Similarity. 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 951–962, Minneapolis, Minnesota. Association for Computational Linguistics.

Metaphors in Literary Machine Translation: Close but no cigar?

Alina Karakanta, Mayra Nas, Aletta G. Dorst
Leiden University Centre for Linguistics
[a.karakanta|a.g.dorst]@hum.leidenuniv.nl

Abstract

The translation of metaphorical language presents a challenge in Natural Language Processing as a result of its complexity and variability in terms of linguistic forms, communicative functions, and cultural embeddedness. This paper investigates the performance of different state-of-the-art Machine Translation (MT) systems and Large Language Models (LLMs) in metaphor translation in literary texts (English→Dutch), examining how metaphorical language is handled by the systems and the types of errors identified by human evaluators. While commercial MT systems perform better in terms of translation quality based on automatic metrics, the human evaluation demonstrates that open-source, literary-adapted NMT systems translate metaphors equally accurately. Still, the accuracy of metaphor translation ranges between 64-80%, with lexical and meaning errors being the most prominent. Our findings indicate that metaphors remain a challenge for MT systems and adaptation to the literary domain is crucial for improving metaphor translation in literary texts.

1 Introduction

In 2015, Toral and Way carried out two landmark studies on Literary Machine Translation (LitMT) challenging 'the perceived wisdom [...] that MT is of no use for the translation of literature' (2015a, p. 123) and the claim that literature remains 'the last bastion of human translation' (p. 123). Despite recent improvements in MT quality, they doubted whether MT would be able to tackle what has been called 'perhaps the most creative task a human translator can take on' (Rothwell et al., 2023, p. 10). Yet Toral and Way (2015a; 2015b) convincingly showed that MT has potential in assisting human literary translators, especially in the translation of fiction novels between closely related languages. Their best-performing system equalled professional human quality almost 20% of the time, and a human evaluation with native speakers indicated that over 60% of the translations were considered of the same or even higher quality. A small but steadily growing number of studies has been conducted in LitMT for different genres and languages (Voigt and Jurafsky, 2012; Besacier, 2014; Thai et al., 2022; Toral et al., 2023; Ploeger et al., 2024), showing significant quality gains of literary-adapted NMT systems over general-purpose MT.

Nevertheless, several challenges remain in LitMT. The time is not yet ripe to admit defeat and concede MT's triumphant victory over the human translator. The gap between LitMT and publishable translations is still large, with MT systems lacking in terms of adequacy, style and tone, and the translation of figurative language (Matusov, 2019; Hansen and Esperança-Rodier, 2022). One characteristic of literary texts that continues to pose difficulties is the use of metaphors, which are problematic for NLP (Chakrabarty et al., 2021) and notoriously hard to translate, even for humans, because of their linguistic and cultural embeddedness.

Recently, Large Language Models (LLMs) have demonstrated remarkable performance in several linguistic tasks, including MT (Kocmi et al., 2024). Unlike traditional encoder-decoder models, LLMs have shown potential in translating long documents, in performing style transfer in a zero-shot manner and have even been tested as aids in creative processes (Chakrabarty et al., 2024). These new abilities lead to the questions: Can LLMs address key challenges in LitMT? How well do they perform on metaphor translation, a hallmark of creative expression? To date, only Dorst (2024) and Zajdel (2022) have studied metaphor in LitMT, but each performed a qualitative analysis on a single text and engine. No studies have systematically com-

Proceedings of Machine Translation Summit XX Volume 1, pages 276–286, June 23-27, 2025

pared how different MT systems translate different types of metaphor and whether LLMs offer new possibilities in addressing this persistent problem.

In this paper, we investigate metaphors in LitMT by analysing (i) the performance of state-of-the-art MT systems in translating metaphor and ii) the kinds of errors the different systems make when translating metaphor in literary texts. Our contributions are as follows:

- We compile a new parallel test set of literary texts and their translations (En→NL), annotated with metaphors at the word-level.[1]
- We conduct an evaluation of several commercial and open-source, encoder-decoder and decoder-only, generic and literary-adapted systems in their performance of translating literary texts from English into Dutch using multiple metrics.
- We complement the automatic evaluation with a human evaluation of the accuracy of the systems in translating metaphors, by annotating the errors in metaphor translation and classifying them based on error type.
- Our findings show that metaphor translation is still a challenge in LitMT and that adaptation to the literary domain (regardless of the model architecture) is crucial for addressing metaphor translation in literary texts.

2 Related work

2.1 Metaphor translation

Since the cognitive turn of the 1980s, metaphors are no longer seen as instances of 'deviant' or 'decorative' language use but recognized as a fundamental cognitive tool in human understanding and communication. Metaphors allow us to think and talk about abstract, complex and unfamiliar concepts (such as time, life or arguments) in terms of more concrete, simple and familiar ones (such as concrete objects, movement or living entities). For example, this is why we said in the Introduction that Toral and Way carried out their 'landmark' studies 'in' 2015 and that their studies 'challenged' the widely believed claim that literature is the 'bastion' of human translation. Lakoff and Johnson's (1999; 2008) groundbreaking work showed that such metaphorical uses of words and phrases form systematic patterns in language because they realize underlying conventional conceptual metaphors in thought. For instance, 'in 2015' realizes TIME IS SPACE and both 'challenged' and 'bastion' are realizations of ARGUMENT IS WAR (where beliefs and theories are locations we defend or attack, gaining or losing ground, challenging our opponents until we win or lose the argument). Since most of the metaphors we use are conventional both in language and thought, we normally use and understand them automatically and effortlessly, without realizing they are metaphors.

Yet even highly conventional linguistic metaphors quickly become problematic once we try to translate them. In fact, metaphors have long been considered a notorious problem in translation as a result of their complexity, variability and linguistic and cultural embeddedness – Newmark even went as far as to consider metaphors "the most important particular problem in translation" (1988, p.104). While a small but consistent stream of studies has focused on detailing procedures for metaphor translation (e.g. Van den Broeck (1981); Newmark (1988); Mandelblit (1995); Dickins (2005); Ali (2006)), most of these focus on metaphor at the linguistic level and finding equivalent forms in the target language (but see Schäffner (2017); Shuttleworth (2017)). Very little attention has been paid to the communicative and rhetorical function of metaphor and the role metaphors play in creating aesthetic effects or stylistic coherence, issues particularly relevant in literary translation where style and content are inseparable (Landers, 2001; Boase-Beier, 2014). As illustrated by Dorst (2019) metaphor translation based on local decisions without considering global textual patterns may disrupt a text's stylistic coherence. A subsequent study by Dorst (2024) on the differences between human and machine translations of literary metaphor found that the human translator frequently opted for deletion and normalization, especially for creative metaphors. When the metaphors were conventional, especially fixed collocations and idiomatic expressions, human translators, both professionals and students, showed more (creative) variation in their solutions and the MT system (Google Translate) made more mistakes (lexical and/or grammatical errors). The current study picks up from this point to investigate more systematically how different MT systems –

[1]While the Dutch translations cannot be released at the time of writing due to pending copyright approval, our annotations of metaphors and code can be found at: https://github.com/fatalinha/MetaphorMT/tree/main.

NMT and LLM – compare in their translation of metaphor and the type of errors they make.

2.2 Literary machine translation

The suitability and feasibility of MT for the literary domain has been a long-standing topic of inquiry in MT research. Techniques that have demonstrated improvements include domain adaptation (Toral and Way, 2015a,b; Toral et al., 2023), author-tailored adaptation (Kuzman et al., 2019; Oliver, 2023), restoration of lexical richness to that of the source text (Ploeger et al., 2024) and automatic post-editing (Thai et al., 2022). Studies assessing MT quality in literary contexts have recognised the importance of conducting human evaluations and error analyses of the generated outputs in addition to computing automatic evaluation metrics. While readers seem to rate a significant percentage of MT sentences as acceptable, error-free or equivalent to human translations, with variations across language pairs (34% for English to Catalan (Toral et al., 2018), ∼20% for English into Russian and German (Matusov, 2019), 44% for English into Dutch (Fonteyne et al., 2020)), a recent multilingual study involving 20 language pairs reported that professional translators preferred human translations in 85% of the cases (Thai et al., 2022). Productivity gains from using MT and post-editing have also been reported as moderate success stories of LitMT (Besacier, 2014; Kuzman et al., 2019). Professional translators, however, still prefer human translation over post-editing for literary texts, mentioning sentence-level fragmentation, wrong level of politeness, vocabulary use, figurative language and cultural items as main limitations of MT (Moorkens et al., 2018).

Another line of research has focused on identifying common error types in LitMT. One notable issue identified is that MT systems often struggle with maintaining referential cohesion (Voigt and Jurafsky, 2012) and have limited potential in addressing the difficulties of literary translation (Jones and Irvine, 2013), mainly because the typical sentence-level MT pipeline is insufficient for this task, as document-level context is critical for the literary domain. Although NMT demonstrated clear improvements in fluency over statistical MT (Toral et al., 2018), adequacy errors and mistranslations are still primary sources of failure (Hansen and Esperança-Rodier, 2022), with fluency aspects such as coherence and style & register still being present (Fonteyne et al., 2020). Discourse-level errors such as coreference and pronoun consistency were identified by Thai et al. (2022), along with overly literal translations. From English to Arabic, translations were found to lack proper handling of idioms and colloquialisms (Omar and Gomaa, 2020). As stated above, the current study builds on the analysis of Dorst (2024), contributing to the ongoing research on the feasibility of LitMT by introducing a previously unexplored aspect, that of metaphor translation.

3 Methodology

3.1 Data

The data for evaluating the models consists of excerpts of four English fiction texts from the VUAMC corpus (Steen et al., 2010) and their published Dutch translations. Since the VUAMC corpus only contains English texts, the Dutch translations were scanned from the physical books, OCR'd and corrected, and manually aligned to the English excerpts at the sentence level. The resulting test set contains 482 sentences (about 6700 words). Details about the selected excerpts can be found in Appendix A.

3.2 Models

The models used in this study were selected to cover a wide range of architectures and system types, both encoder-decoder and decoder-only language models, open- and closed-source, generic and literary-adapted. The selection was guided by the best-performing systems in the literary domain from WMT 2024 (Kocmi et al., 2024). Specifically, the models tested were the following:

1. Commercial NMT systems: Google Translate[2], DeepL[3] and ModernMT[4];

2. S3Big: a literary-adapted NMT Transformer model trained using Marian[5] on general-domain and back-translated literary monolingual data, and then fine-tuned on real in-domain data (parallel novels En→Nl). This is a sentence-level model (Toral et al., 2023);

3. General purpose LLMs: GPT4, GPT4o, and GEITje 7B Ultra (Vanroy, 2024), a conversational LLM fine-tuned for Dutch, based on

[2]https://translate.google.com/
[3]https://www.deepl.com/
[4]https://www.modernmt.com/
[5]https://marian-nmt.github.io/

Mistral and aligned with AI feedback via Direct Preference Optimisation; and

4. Translation-adapted LLMs: Tower-Instruct-7B-V0.2 and 13B-V0.1 (Alves et al., 2024). These language models have been trained on multilingual data and fine-tuned on translation-specific data, so as to handle several translation-related tasks, e.g. translation, paraphrasing, automatic post-editing.

LLMs received a simple prompt in the form: "Translate the following sentence from English into Dutch (NL)". GPT4 and GPT4o were accessed through the Trados Studio OpenAI API on 23 July 2024, with temperature set to 0.75. To test whether prompting can have a positive effect on translation quality in LitMT, a literary prompt was also tested with GPT4o and Tower13b, mentioned here as GPT4o-Lit and Tower13b-Lit respectively: "You are a professional translator, specializing in the translation of literary texts. Translate the following sentence from an English novel into Dutch (NL), paying special attention to the translation of metaphors". For the TowerInstruct models, the ChatML prompt templates format was used. The models were tested with the default settings and batch size 256.

3.3 Evaluation

The system outputs were evaluated for general translation quality against the human reference using multiple automatic metrics: SacreBLEU (Post, 2018), COMET (Rei et al., 2020), BERTScore (Zhang* et al., 2020) and MetricX (Juraska et al., 2023) [6]. MetricX is a learned regression-based metric based on the mT5-XXL pretrained language model. It achieved among the highest correlation with human judgements in the WMT 2024 Metrics task (Freitag et al., 2024). We use MetricX-24-Hybrid-Large and the corpus-level score is computed by averaging the segment-level scores. BERTScore was computed using the MATEO framework (Vanroy et al., 2023). The selection of string-based, neural and LLM-based metrics aims to compare the rankings assigned to the systems by different evaluation metrics and examine the relation between various types of metrics and the quality of metaphor translation.

In addition, the systems' accuracy in translating different types of metaphors was assessed via human evaluation. One hundred sentences were randomly selected from the test set (4 chunks of approx. 6 sentences from each novel) containing 333 source metaphors in total. The outputs of the five highest-performing systems according to the automatic metrics, as well as the official translations, were annotated in INCEpTION (Klie et al., 2018) by two professional translators, native speakers of Dutch. The outputs were presented to the annotators without any information about which system generated which sentence. For each metaphor in the source (annotated at the word level following VUAMC), the evaluators had to detect the corresponding translation in Dutch and assess whether the translation was "correct" or "incorrect". Subsequently, the errors were analysed and classified in three categories: meaning errors (the Dutch translation of the source metaphor has the wrong meaning), form errors (the Dutch translation of the source metaphor is ungrammatical or unidiomatic) or omissions, when the metaphor is left out in the translation.

Following VUAMC, the metaphor translations were also annotated at the word level. However, this word-based approach is not without problems, since words in the source may be expressed by a multi-word expression in the target and vice versa. For example, the English verb 'glare' is correctly translated into Dutch as 'boos kijken' [lit. 'angry look'] while 'wiped out' translates as 'vernietigd' ['destroyed']. In addition, metaphors frequently form multi-word expressions (MWE) (e.g. collocations, idiomatic expressions) in which the translation of the metaphorical word may be considered correct in isolation but not in the multi-word expression. For example, in the phrase 'made good time' the verb 'made' is annotated as a source metaphor and in isolation the translation 'maakte' is technically correct, but the combination 'maakte goede tijd' is incorrect because it is ungrammatical and unidiomatic. An additional problem is that in some cases it is clear that there is an error, because the MWE as a whole is incorrect in the Dutch translation, but it is hard to pinpoint which individual word(s) to annotate. Despite this, the word-based approach is necessary to obtain a measure of accuracy in metaphor translation.

After collecting the error annotations, the per-

[6] SacreBLEU:nrefs:1|bs:1000|s:12345|c:mixed|eff:no|tok:13a|smooth:exp|v:2.4.3
BERTScore: nrefs:1|bs:1000|s:12345|l:other|v:0.3.12|mateo:1.1.3
COMET: nrefs:1|bs:1000|s:12345|c:Unbabel/wmt22-comet-da|v:2.2.2

centage of correctly translated metaphors is reported per system. We compute inter-annotator agreement using Cohen's κ (Cohen, 1960), based on whether annotators agree on their judgement of a source metaphor being translated correctly or not. In addition, we report inter-rater reliability (IRR) as the percentage of matches between the two raters.

4 Results

4.1 Automatic evaluation

Table 1 presents the automatic quality scores for the various systems. In response to our first research question "Which is the highest-performing system for LitMT from English into Dutch?", notably, **different types of metrics assign higher scores to different systems, not allowing to pinpoint a clear winner**. The best-performing system based on the neural metrics is the commercial system Google Translate (GT) with a COMET score of 84.02 and a BERTScore of 83.96, while BLEU favours DeepL with a score of 31.31 and 3 points difference from GT, the second-scoring system. However, MetricX scores the output of GPT4o-Lit as the best with a score of 2.0461 (the lower the score the better). Similar to the neural metrics, MetricX scores Google Translate higher than DeepL with a score of 2.1075 and 2.1545 respectively. It appears that MetricX favours the outputs of GPT models, but not those of the Tower models, even though all of them are LLMs. These differences in ranking highlight variations in how each metric evaluates translation quality and the difficulty of relying solely on automatic evaluation in LitMT.

Another hypothesis put forward in the Introduction is that LLMs may outperform NMT systems in LitMT. However, based on the automatic scores, **there is no clear indication that LLMs can surpass NMT systems yet**. LLMs perform similarly with commercial systems and the literary-adapted system S3Big. Bootstrap resampling on COMET and BERTScore scores shows a second-place tie among DeepL, the GPT4 models, and S3Big (light gray). A similar second tie is observed for BLEU scores. This is notable, given that GPT models have not been explicitly trained for translation or on literary data. On the contrary, translation-specific LLMs (Tower7b and 13b) unexpectedly scored significantly lower according to all metrics, forming a third tie together with ModernMT. Lastly, GEITje has the lowest score, despite being fine-tuned for Dutch. This is expected since it is not fine-tuned to the task of translation, which often leads to hallucinations and the inability to follow instructions. Therefore, vanilla LLMs do not seem to be bringing a transformative breakthrough in LitMT yet.

LLM adaptation to translation tasks or the target language did not demonstrate promising results, but does adaptation to the literary domain make a difference? S3Big performs on par with commercial systems, **demonstrating that domain adaptation can still yield high-performing non-commercial NMT systems**. Similar results were reported by Toral et al. (2023) where S3Big showed only a 2% reduction in COMET compared to DeepL. Even though MetricX assigned a low score to S3Big, the best score was assigned to GPT4o-Lit, the system adapted with the literary prompt. For the neural and string-based metrics, the literary prompt (GPT4o-Lit) also led to minor improvements in scores. However, this is not the case with Tower13b, where the literary prompt hurt performance. Both these observations indicate that further adaptation and careful fine-tuning of LLMs to the literary domain could lead to improvements in LitMT, a direction to be explored in future work.

To sum up, the automatic evaluation suggests that commercial NMT systems are the strongest for LitMT, followed by closed-source LLMs. However, the open-source, literature-adapted NMT system S3Big remains competitive despite being trained on significantly less data, demonstrating the effectiveness of domain adaptation. In contrast, open-source LLMs still lag behind, even when specifically trained for translation tasks. However, another question remains: How accurate are these systems in translating metaphors? To answer this question, the top performing systems from each architecture group are selected to conduct a human evaluation of their accuracy in metaphor translation. The selected systems include DeepL, Google Translate (GT), GPT4, Tower13b and S3Big.

4.2 Human evaluation

Table 2 shows the human evaluation scores in metaphor translation for the selected systems, as well as for the human translation (Ref). In general, the scores for the accuracy of translating metaphors range between 64-80%, showing that **metaphors are still a challenge for MT systems**. The literary-adapted NMT system S3Big has the highest accuracy in translating metaphors with 75% of the metaphors on average annotated as correctly trans-

system	MetricX24↓	COMET↑	BERTScore↑	BLEU↑
DeepL	2.1545	83.55	83.46	**31.31**
Google Translate	2.1075	**84.02**	**83.96**	28.47
ModernMT	2.4435	82.89	82.83	26.30
GPT4	2.1464	83.45	83.42	27.59
GPT4o	2.1256	83.18	83.15	26.35
GPT4o-Lit	**2.0461**	83.25	83.20	26.68
GEITje	3.7430	77.64	77.62	14.89
TOWER7b	2.3195	82.73	82.72	23.66
TOWER13b	2.2156	82.81	82.83	24.53
TOWER13b-Lit	2.3778	82.04	82.13	24.94
S3Big	2.3593	83.31	83.30	28.72

Table 1: MetricX24, COMET, BERTScore and BLEU scores of different systems on the En→Nl literary test set. Best score in **bold**. Different colours (light blue , medium blue and dark blue) indicate statistically significant differences between systems. Systems sharing the same colour are not statistically different from each other.

	Ref	DeepL	Google Tr.	GPT4	Tower13b	S3Big
An1	87%	78%	76%	75%	75%	**80%**
An2	87%	**70%**	68%	65%	64%	**70%**
Avg	87%	74%	72%	70%	69.5%	**75%**
κ	0.57	0.47	0.52	0.47	0.46	0.42
IRR	90%	79%	80%	77%	77%	78%

Table 2: Accuracy in the translation of metaphors by the two annotators (An1 and An2) and on average (Avg), Cohen's κ and inter-rater reliability (IRR). Different colours (light blue , medium blue and dark blue) indicate statistically significant differences between systems ($p < .05$) based on pairwise comparisons. Systems sharing the same letter are not significantly different from each other.

lated. The second-best system was found to be DeepL (74%). The highest-performing system based on the neural metrics, GT, comes third (72%). However, a logistic mixed-effects model did not reveal statistically significant differences in accuracy between S3Big, DeepL and GT. LLMs, despite promises to address issues in LitMT, have the lowest scores in metaphor translation with 70% for GPT4 (β = -0.339, p = .019 compared to S3Big) and 69.5% for Tower13b (β = -0.398, p = .006), suggesting that adaptation techniques may be required for these systems to address literary aspects more accurately.

Interestingly, metaphors in the human translation were also sometimes annotated as incorrect (84% accuracy). Most of the identified errors in the human translation were meaning errors or omissions (see Table 3 for a classification of errors). These appeared to occur when the source metaphors were rather hard to interpret or their meaning was ambiguous (for example, 'knotted' in 'once free of the knotted tentacles of the suburbs') or when the translation may have sounded forced or awkward rather than literary and creative. In such cases, the human translator may have decided to go for the "safe" option of omitting the metaphor. After all, as pointed out by Guerberof-Arenas and Toral (2022), creativity involves both novelty and acceptability. This is a particularly interesting area for future investigations: while omission is generally considered an error in MT, it is often a deliberate risk-avoiding strategy in human translation. Future studies could explore in more detail whether metaphor omissions in MT occur in the same contexts and under the same conditions as in HT.

On the total number of annotations, a moderate agreement was found between the annotators with Cohen's κ at 0.49 (Landis and Koch, 1977) and a total IRR of 80%. The annotators agreed more on their assessments of the human translation (κ=0.57, IRR=90%) and less on the metaphors translated by S3Big (κ=0.42, IRR=78%). When comparing the scores of the two annotators, we observe that Annotator 2 was more strict than Annotator 1 when

assessing the machine-translated metaphors, by 9% on average, even though the annotators agree on the percentage of correct human-translated metaphors. The moderate inter-annotator agreement shows that the task of assessing metaphor translations is difficult and even trained professional translators may disagree on whether a particular metaphor translation counts as an error. As discussed above, this may be due to the inherent difficulty of pinpointing whether and where errors occur in metaphor translation. Similar agreement scores have been reported in other studies involving error annotation in literary translation (Fonteyne et al., 2020), showing a potential subjectiveness of error assessment in the literary domain. More importantly, what professional translators or linguists consider errors may at times be considered acceptable or creative by the average reader, especially in literary texts.

5 How do MT systems translate metaphors?

The automatic evaluation (Table 1) indicated that the commercial NMT systems obtained the highest quality overall for literary MT. The human evaluation (Table 2) showed that the literary-adapted NMT system S3Big had the highest accuracy for metaphor translation, together with DeepLand GT, with LLMs falling short. To determine whether the NMT and LLM systems make the same or different types of errors and compare the types of errors with the human translation, Annotator 1 labelled the identified errors for their error type, i.e. Form, Meaning or Omission. Table 3 shows the error types by system. A total of 228 errors were identified in all outputs and the human translation by Annotator 1, divided in 128 meaning errors, 85 form errors and 15 omissions. In general, meaning errors are the most prevalent in all MT systems, however, there are differences: form errors are more common for GT while for LLMs (GPT and Tower) as well as for the literary-adapted system S3Big the difference between the number of meaning errors and form errors is much clearer.

Overall, the observations made during the error annotation support previous findings that lexical errors (mistranslations) are the most frequent type of error. Conversely, this raises the question whether lexical errors are often the most frequent type of error in MT output because of the pervasiveness of metaphors in everyday discourse. A closer look at the translations shows that, as suggested by Dorst (2019, 2024), most of these meaning errors concern highly conventional linguistic metaphors. For example, in 'the dark *mouth* of a concrete pillbox' (S10, C8T), 'mouth' was incorrectly translated by DeepL and GPT as 'mond' [mouth] and by Google Translate as 'monding' [mouth, estuary]; Tower uses the correct 'opening' [opening] and S3Big system the correct 'ingang' [entrance]. Similarly, in '*Wiped out* twenty million Russians' (S132, G0L), Google Translate has the incorrect 'weggevaagd' [erased, swept away], GPT the incorrect 'verwijderd' [removed] and S3Big the incorrect 'uitgebuit' [exploited], while DeepL and Tower use the correct 'uitgeroeid' [exterminated]. Table 3 shows that the human translator made the fewest meaning errors, and the two commercial systems slightly fewer than the LLMs and S3Big.

For the form errors, the situation is slightly different: DeepL and S3Big make fewer form errors than the other systems. Here, the contrast is quite striking between Google Translate (with 22 form errors) and S3Big (with only 9 form errors). This suggests that DeepL and S3Big may be better at correctly translating multi-word metaphors such as idiomatic expressions. For example, both systems correctly translated '*keep* your voice *down*' as 'praat niet zo hard' [lit. talk not so loud] rather than the incorrect (unidiomatic) direct translation 'houd je stem laag' produced by Google Translate and GPT. The translation produced by Tower - 'houd je stem maar eens in' - is particularly puzzling because it sounds idiomatic but is in fact meaningless, since 'inhouden' is something you can do with your breath (e.g. hold your breath) but not with your voice. The combination 'je stem inhouden' simply does not exist in Dutch (but 'je adem inhouden' does). Something slightly different happened with the expression 'get under his skin' (S89, G0L), where DeepL, Google Translate and GPT all have the incorrect direct translation 'onder zijn huid kruipen' (which does not exist as a conventional metaphorical expression in Dutch), while Tower outputs a correct and idiomatic alternative 'van streek te maken' [to upset] and S3Big outputs the idiomatic but incorrect (wrong meaning) 'overdonderen' [= overwhelm].

More examples need to be collected and analysed to determine whether these patterns are consistent across larger datasets but this first exploration may suggest that as NMT and LLMs become better at avoiding incorrect direct translations of multi-word expressions they may start making more meaning errors (which will be harder

System	Meaning	Form	Omission	Total
Ref	10 (50%)	3 (15%)	7 (35%)	20 (100%)
DeepL	18 (49%)	14 (38%)	5 (13%)	37 (100%)
GT	20 (48%)	22 (52%)	0 (0%)	42 (100%)
GPT4	25 (57%)	19 (43%)	0 (0%)	44 (100%)
Tower13b	30 (61%)	18 (37%)	1 (2%)	49 (100%)
S3Big	25 (70%)	9 (25%)	2 (5%)	36 (100%)
Total	128 (100%)	85 (100%)	15 (100%)	228 (100%)

Table 3: Errors in meaning form and omissions for each system and reference translation.

to spot, especially for readers without access to the source text). LitMT is advancing to a stage in which the number of obviously incorrect and unidiomatic translations is decreasing, and some of the metaphor translations are indistinguishable from human translation. However, the big question is whether the remaining errors - both form and meaning - affect the readers' understanding of the metaphors and their role in the narrative. If the cost for obtaining idiomatic metaphor translations is a shift in meaning is that a price we are willing to pay?

6 Conclusion and future work

In this paper, we addressed the translation of metaphor in literary MT from English into Dutch by comparing different transformer-based architectures. We investigated how the different systems translate metaphors and determined what type of errors they tend to make, asking whether LLMs provide new opportunities in tackling this long-standing challenge in NLP. Regarding the performance of state-of-the-art MT systems in metaphor translation, our conclusion is that they come close, but no cigar. The automatic evaluation showed that different types of metrics favoured different systems and no single system consistently outperformed the others. No clear evidence was found indicating that LLMs in their current setting outperform NMT systems in LitMT and metaphor translation, at least for this language pair and type of literary content. Commercial NMT systems produced the overall highest quality output, followed by closed-source LLMs. Notably, the open-source, literature-adapted NMT system S3Big remained competitive despite having been trained on significantly less data, demonstrating the effectiveness of domain adaptation. Additionally, the human evaluation revealed that the accuracy of the systems for metaphor translation specifically ranged between 64-80%, highlighting that metaphors remain a challenge for MT systems. A closer look at the errors identified in human evaluation revealed that most were meaning errors (i.e. lexical) rather than form errors (i.e. grammatical) and most of the errors concerned highly conventional metaphorical expressions.

Our current findings suggest that further research is needed to assess whether the errors made by MT - both in form and meaning - affect readers' understanding of the metaphors and the narrative. In addition, the structural similarities between English and Dutch may result in more "false friend" metaphor translations, which may appear to be fluent and correct while the metaphor technically does not exist in Dutch. The next phase of this project is therefore to also extend it to more languages. Another continuing objective of the current project is to develop a clearer taxonomy to identify and label different types of errors and shifts in metaphor translation, especially given the difficulties in deciding where (which word) the error occurs, what type of error it is, and whether it should be counted as an error or as a creative solution. We are therefore also conducting a reader-response study that investigates how readers respond to the different MT metaphor translations that were classified as errors in our human evaluation.

Sustainability statement

The experiments presented in this paper involving running model inference and computing neural automatic metrics ran for 7h and 30min on 1 GPU NVIDIA A100 40GB PCIe, while larger models ran for 1h on 1 GPU NVIDIA A100 80GB PCIe. In total, our experiments drew 1,389 kgCO2e. Based in [country removed for anonymity], this had a carbon footprint of 3.70 kWh, which is equiva-

lent to 1.52 tree-months. (calculated using green-algorithms.org v3.0 (Lannelongue et al., 2021)).

Limitations

In this paper, we addressed the translation of metaphors in literary MT. However, this does not encompass the translation of metaphors in other domains and genres. We evaluated different systems in a high-resource language pair that consists of relatively similar languages, in one translation direction. Given the sparsity of metaphor-annotated data and difficulty of obtaining literary translations we found this difficult to avoid. Experiments with more languages and literary texts may retrieve richer results. In addition, even though our data is transparent, in the sense that we have reported the exact excerpts and sentence numbers from the VUAMC corpus, for the time being we cannot distribute the Dutch translations ourselves, due to copyright restrictions. Lastly, we acknowledge that the list of models, prompts and settings tested is not exhaustive, even though we found it to be representative of the range of models currently available.

References

Abdul Sahib Mehdi Ali. 2006. On the translation of metaphor: Notions and pedagogical implications. *IJAES*, 7:121–136.

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.

Laurent Besacier. 2014. Machine translation for litterature: a pilot study (traduction automatisée d'une oeuvre littéraire: une étude pilote) [in French]. In *Proceedings of TALN 2014 (Volume 2: Short Papers)*, pages 389–394, Marseille, France. Association pour le Traitement Automatique des Langues.

Jean Boase-Beier. 2014. *Stylistic approaches to translation*. Routledge.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024. Creativity support in the age of large language models: An empirical study involving professional writers. In *Proceedings of the 16th Conference on Creativity & Cognition*, C&C '24, page 132–155, New York, NY, USA. Association for Computing Machinery.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online. Association for Computational Linguistics.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:1–37.

James Dickins. 2005. Two models for metaphor translation. *Target*, 17(2).

Aletta G Dorst. 2019. Translating metaphorical mind style: Machinery and ice metaphors in ken kesey's one flew over the cuckoo's nest. *Perspectives*, 27(6):875–889.

Aletta G. Dorst. 2024. Metaphor in literary machine translation: style, creativity and literariness. In *Computer-Assisted Literary Translation*, pages 173–186. New York: Routlegde.

Margot Fonteyne, Arda Tezcan, and Lieve Macken. 2020. Literary machine translation under the magnifying glass: Assessing the quality of an NMT-translated detective novel on document level. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3790–3798, Marseille, France. European Language Resources Association.

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.

Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*, 11(2):184–212.

Damien Hansen and Emmanuelle Esperança-Rodier. 2022. Human-Adapted MT for Literary Texts: Reality or Fantasy? In *Proceedings of the New Trends in Translation and Technology Conference*, pages 178–190, Rhodes, Greece.

Ruth Jones and Ann Irvine. 2013. The (un)faithful machine translator. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–101, Sofia, Bulgaria. 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.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

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.

Taja Kuzman, Špela Vintar, and Mihael Arčan. 2019. Neural machine translation of literary texts from English to Slovene. In *Proceedings of the Qualities of Literary Machine Translation*, pages 1–9, Dublin, Ireland. European Association for Machine Translation.

George Lakoff and Mark Johnson. 1999. Philosophy in the flesh—the embodied mindand its challenge to western thought. *NY: Basic Books*.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Clifford E Landers. 2001. Literary translation: A practical guide. *Multilingual Matters*.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green algorithms: Quantifying the carbon footprint of computation. *Adv. Sci.*, 1.

Nili Mandelblit. 1995. The cognitive view of metaphor and its implications for translation theory. *Translation and meaning*, 3(1):483–495.

Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.

Joss Moorkens, Antonio Toral, Sheila Castilho, and Andy Way. 2018. Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7:240–262.

Peter Newmark. 1988. *A textbook of translation*. Prentice Hall International.

Antoni Oliver. 2023. Author-tailored neural machine translation systems for literary works. In *Computer-Assisted Literary Translation*, pages 126–141. Routledge.

A. Omar and Y. Gomaa. 2020. The machine translation of literature: Implications for translation pedagogy. *International Journal of Emerging Technologies in Learning (iJET)*, 15:228–235.

Esther Ploeger, Huiyuan Lai, Rik Van Noord, and Antonio Toral. 2024. Towards tailored recovery of lexical diversity in literary machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 286–299, Sheffield, UK. European Association for Machine Translation (EAMT).

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.

Andrew Rothwell, Andy Way, and Roy Youdale. 2023. *Computer-Assisted Literary Translation (1st ed.)*. Routledge.

Cristina Schäffner. 2017. Metaphor in translation. In E. Semino and Z. Demjen, editors, *The Routledge Handbook of Metaphor and Language*, pages 247–262. Abingdon: Routledge.

Mark Shuttleworth. 2017. *Studying scientific metaphor in translation*. Routledge.

G.J. Steen, A.G. Dorst, J.B. Herrmann, A.A. Kaal, T. Krennmayr, and T. Pasma. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*. Amsterdam: John Benjamins.

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, Andreas Van Cranenburgh, and Tia Nutters. 2023. Literary-adapted machine translation in a well-resourced language pair. explorations with more data and wider contexts. In Andrew Rothwell, Andy Way, and Roy Youdale, editors, *Computer-Assisted Literary Translation*. Routledge: New York.

Antonio Toral and Andy Way. 2015a. Machine-assisted translation of literary text: A case study. *Translation Spaces*, 4:240–267.

Antonio Toral and Andy Way. 2015b. Translating literary text between related languages using SMT. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 123–132, Denver, Colorado, USA. Association for Computational Linguistics.

Antonio Toral, Martijn Wieling, Sheila Castilho, Joss Moorkens, and Andy Way. 2018. Project PiPeNovel: Pilot on post-editing novels. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, page 385, Alicante, Spain.

Raymond Van den Broeck. 1981. The limits of translatability exemplified by metaphor translation. *Poetics today*, 2(4):73–87.

Bram Vanroy. 2024. Geitje 7b ultra: A conversational model for dutch. *Preprint*, arXiv:2412.04092.

Bram Vanroy, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine Translation Evaluation Online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500. European Association for Machine Translation (EAMT).

Rob Voigt and Dan Jurafsky. 2012. Towards a literary machine translation: The role of referential cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 18–25, Montréal, Canada. Association for Computational Linguistics.

Alicja Zajdel. 2022. Catching the meaning of words: Can google translate convey metaphor? In *Using Technologies for Creative-Text Translation*, pages 116–138. Routledge.

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*.

A Dataset

Metaphor in Fiction sample from VUAMC. The following excerpts have been selected as the test set for this study:

- C8T: *Devices and desires*. James, P D (1989). (sentences 2-14, 27-49, 114-131)
- CDB: *A fatal inversion*. Vine, Barbara (1987). (fragment 02: sentences 380-400, 422-465, fragment 04: 855-881)
- FPB: *Crimson*. Conran, Shirley (1992). (1060-1102, 1249-1290, 1312-1373)
- G0L: *The Lucy ghosts*. Shah, Eddy (1993). (sentences 19-69, 75-152, 162-222)

And their Dutch translations:

- PD James – Melodie des doods
- Ruth Rendell – Ongewenst weerzien
- Shirley Conran - Karmozijn
- Eddy Shah – Het Lucy komplot

Synthetic Fluency: Hallucinations, Confabulations, and the Creation of Irish Words in LLM-Generated Translations

Sheila Castilho, *__Zoe Fitzsimmons,__ *__Claire Holton__ and *__Aoife Mc Donagh__
SALIS, ADAPT Centre, Dublin City University
sheila.castilho@dcu.ie, *first.second@mail.dcu.ie

Abstract

This study examines hallucinations in Large Language Model (LLM) translations into Irish, specifically focusing on instances where the models generate novel, non-existent words. We classify these hallucinations within verb and noun categories, analyse whether these hallucinations adhere to Irish morphological rules and what linguistic tendencies they exhibit. Beyond classification, the discussion raises speculative questions about the implications of these hallucinations for the Irish language. Our findings offer food for thought regarding the increasing use of LLMs and their potential role in shaping Irish vocabulary and linguistic evolution. We aim to prompt discussion on how such technologies might influence language over time, particularly in the context of low-resource, morphologically rich languages.

1 Introduction

Since the emergence of neural machine translation (MT), hallucinations have been recognised as a significant challenge in the field (Koehn and Knowles, 2017). LLMs also hallucinate, and despite efforts to mitigate this, hallucinations remain common—especially in low-resource settings (Sennrich et al., 2024). Hallucinations produced by LLMs are claimed to be "qualitatively different from those of conventional translation models" which include "off target translations, overgeneration, and even failed attempts to translate" (Guerreiro et al., 2023, p. 1501).

This study focuses on specific types of hallucinations, namely instances where the system *invents* new words during translation. Our goals are to identify which word classes are affected when open LLMs generate hallucinations in a low-resource language like Irish (Gaeilge) and to assess whether these hallucinated words follow Irish linguistic rules or diverge from established conventions. To the best of our knowledge, this is the first study to examine the morphology of hallucinations generated when translating into Irish.

2 Background

2.1 Hallucinations

Recent advancements in artificial intelligence, particularly the rise of decoder-only LLMs like GPT-3.5 and GPT-4, have ushered in a new era for MT (Brown et al., 2020; Hendy et al., 2023; Moslem et al., 2023).

Despite their impressive capabilities, generative LLMs continue to face significant difficulties when translating low-resource languages (Castilho et al., 2023; Robinson et al., 2023). These challenges arise from the severe under-representation of low-resource languages in available training data. As a result, the translations produced often reflect "poor generalization" and may be "inaccurate or nonsensical" due to the models' "limited exposure to the linguistic nuances" of these languages (Shu et al., 2024). One of the issues LLMs face is that of hallucinations (Bang et al., 2023). Several works have recorded the types of hallucination that LLMs produce in different NLP tasks (Ji et al., 2023; Huang et al., 2024). In the context of narrative and dialogue generation, Sui et al. (2024) suggest that hallucination are not necessarily "inherently harmful" and may offer potential benefits, referring to them as "confabulations."

Few studies have addressed hallucination in LLM-based MT (Guerreiro et al., 2023; Sennrich et al., 2024; Briakou et al., 2024), with most focusing on the detection and mitigation. Our interest lies in examining the morphology of specific types of hallucinations and confabulations, particularly those involving the creation of entirely new words by the systems.

*These authors contributed equally to this paper and are named alphabetically.

Proceedings of Machine Translation Summit XX Volume 1, pages 287–299, June 23-27, 2025

2.2 Irish Morphology

Examining the morphology of invented hallucinations is especially relevant for the Irish language, as its morphological structure relies heavily on the use of suffixes, infixes and prefixes (Cassidy, 2024). Like other morphologically rich languages, Irish exhibits a high degree of inflection and relatively free word order, which poses specific challenges for MT from English (Cotterell et al., 2018). These structural differences often lead to errors in translation output, particularly when models trained predominantly on English struggle to accurately generate complex morphological forms or correctly interpret flexible syntactic structures (Lankford et al., 2021). Cotterell et al. (2018) suggest that morphological typology may explain some of the variability in model performance across languages, noting that LLMs tend to perform worse on highly inflected languages. Arnett and Bergen (2025) highlight performance gaps across morphological types, raising concerns about linguistic disparities in NLP. Given these findings, investigating hallucinations in Irish—a morphologically rich but low-resource language—may provide insights into broader trends affecting similar languages.

Irish **nouns** are categorised into five declensions (Appendix A), where "the defining criterion [...] is the form of the genitive singular ending" (Ball and Muller, 2010, p.177). The construction of the plural form in Irish consist of two categories, '*lagiolraí*' (weak plurals) and the '*treaniolraí*' (strong plurals). Weak plurals are mainly found in the first and second declensions. In the first declension, plural formation typically involves palatalisation of the final consonant, whereas in the second declension, many nouns form their plurals by adding -a to the singular form. Strong plurals encompass all other plural formations, as nouns in the 3rd, 4th and 5th declensions take strong plural endings. Examples of such endings include *-(e)acha, -(e)annna, -(a)í, -t(h)a and -t(h)e*.

Verbs have the addition of initial mutations, such as lenitions and eclipses. Each tense and mood has its own unique set of suffixes for the conjugation of verbs (see Appendix B). Irish verbs are formed by classification into two conjugations: the first and second conjugations (*an chéad réimniú agus an dara réimniú*). The first conjugation consists of "all one-syllable verbs, two-syllable verbs ending in *-(e)áil* and a small number of two-syllable verbs, which are not syncopated (lose their second syllable) when a third or fourth syllable is added" (Ball and Muller, 2010, p.189). The second conjugation is comprised of all other two-syllable verbs. Within the first and second conjugations, there are two possible suffixes depending on the type of vowels in the roots. **Broad vowels** (*leathan*) *-a, -o, -u* must be followed by the suffix beginning with a broad vowel; and **Slender vowels** (*caol*) *-i, -e* must be followed by the suffix beginning with a slender vowel. A *lenition* marks the past and imperfect tenses, the conditional mood and also follows the negative particles, the conjunction *má* and the interrogative particle *ar*, and is also used following the direct relative clause particle *-a*.[1]

Since the seventeenth century, Irish has been influenced almost entirely by English with "the most dramatic changes have occurred in the last 100 years, in the period when the monolingual Irish speaker became a rarity" (Hickey, 2009, p. 671). As such, there is a tendency to borrow lexicon from English, and adapt these borrowings to align with its grammatical and morphological rules, known as lexical borrowing with adaptation (Mulhall, 2018). Similarly, "new loans replacing existing Irish words", which has been referred to as 'detrimental change' has been noted in recent years (Hickey, 2009, 671). One example is the word 'zoo', which appears in de Bhaldraithe's 1959 English-Irish Dictionary as *gairdín ainmhithe* (garden of animals).[2] In Ó Dónaill's 1977 *Foclóir Gaeilge-Béarla* (Irish-English Dictionary)[3] the word 'zoo' appears as *zú*, while the previous translation is no longer listed. In parallel, code-switching—defined as "instances of the linguistic phenomenon that results in mixed-language text" (Lynn and Scannell, 2019, p.33) — has also become increasingly common in contemporary Irish usage (Cassidy, 2024).

2.3 Automatic Translation of Irish

Due to its intricate morphology described above, not to mention the rich inflectional system, the Irish language poses significant challenges when translating from English. These challenges are even more pronounced for automatic systems, where maintaining grammatical accuracy in features such as noun gender and case inflections proves particularly difficult (Lankford et al., 2023). Nonetheless, the challenge of MT for Irish has been documented

[1]Table 16 in the Appendix, shows an example of the four possible categories for suffixes when conjugating verbs.
[2]https://www.teanglann.ie/ga/eid/zoo
[3]https://www.teanglann.ie/ga/fgb/zoo

in several research works (Dowling et al., 2018, 2020; Lankford et al., 2021).

Regarding LLMs for Irish, we highlight the work of Lankford et al. (2023) fine-tuned multilingual models for translating low-resource languages, including Irish. Tran et al. (2024b) report on an effort to develop an open-source Irish-based LLM. Their results demonstrate strong performance in both understanding and generating Irish text; however, issues such as the forgetting of English as "a consequence of continued pre-training on Irish data" remain (Tran et al., 2024a, p.194).

For a better understanding of the models' ability to handle these complexities, we analyse whether hallucinated words generated by LLMs conform to Irish morphological rules or diverge entirely. We draw on the definitions of hallucinations proposed by Huang et al. (2024), who classify them into *factuality hallucination* and *faithfulness hallucination*, where the former is a discrepancy on verifiable real-world fact, and the latter "captures the divergence of generated content from user input or the lack of self-consistency within the generated content" (ibid, p.42:2). Moreover, *faithfulness hallucination* is subdivided into *context inconsistencies*, which arise when generated content misaligns with the provided context.[4] Under this definition, the phenomenon of the model inventing new words falls within the category of faithfulness hallucinations, specifically as context inconsistencies.

The term "*confabulation*" has been proposed as a more accurate alternative to hallucination. Sui et al. (2024) argue that "LLM confabulations mirror a human propensity to utilize increased narrativity as a cognitive resource for sense-making and communication" (p.14274). They define confabulation as a narrative-driven tendency to organise available information into coherent stories, even when key details are missing—leading to the generation of plausible yet fictional content. From this viewpoint, the model's invention of new words that resemble legitimate Irish morphology can be framed as confabulations. Therefore, in this paper, we use *hallucination* as a general term to refer to all outputs that diverge from the source content or expected translation, while we reserve the term *confabulation* for hallucinated outputs that invent new words which appear internally coherent and plausible according to Irish morphological rules. These definitions provide a foundation for analysing the morphological patterns observed in LLM-generated outputs when translating from English into Irish, allowing us to distinguish between different types of invented words and their potential implications.

[4]"Instruction inconsistency" and "logical inconsistency" are not relevant to the hallucination type studied here.

3 Experimental Setup

3.1 Test Set

Document Title	Domain	# Sentences	# Tokens
Giant fans of wind energy	News	55	4898
Arm processors	News	40	5691
Creating synthetic life	TED Talk	130	13605

Table 1: Test Set Statistics.

To evaluate these types of hallucinations, we conducted a preliminary pilot test, which identified that general texts (such as general news) did not produce any of these hallucinations. However, domain-specific texts, particularly those in scientific and medical fields containing a higher frequency of unfamiliar terms, showed noticeable examples of these hallucinations. Therefore, we selected three texts from the DELA corpus (Castilho et al., 2021)[5] for this experiment: two scientific news texts and one technical TED Talk, as shown in table 1. We note that As our test sets predate 2022, models may have seen them during training. However, this is not problematic, as the terminology was deliberately selected to provoke hallucinations, aligning with our aim to assess model performance on challenging, domain-specific Irish content.

3.2 LLMs

The pilot phase involved testing three open LLMs: ChatGPT 4.0,[6] Co-Pilot, and Gemini.[7] However, both Co-Pilot and Gemini presented significant challenges, as their outputs were notably verbose (Briakou et al., 2024), even after multiple attempts to refine the translation process, with many refusals to translate. Due to these limitations, we decided to focus on two versions of ChatGPT: 4.o (henceforth, GPT4) and 4.o Mini (henceforth, Mini). It should be noted that users accessing ChatGPT 4.o are switched to the Mini version after exceeding the limit of 50 messages within a 3-hour period.[8]

3.2.1 Prompts

Mizrahi et al. (2024, p.935) warn against the limitations of single-instruction evaluation of LLMs,

[5]https://github.com/SheilaCastilho/DELA-Project

[6]https://chatgpt.com

[7]see copilot.microsoft.com and gemini.google.com

[8]https://help.openai.com/en/articles/9275245-using-chatgpt-s-free-tier-faq?utm_source=chatgpt.com

claiming that "a simple rephrasing of the instruction, template can lead to drastic changes in both absolute and relative model performance". We note however that, since our goal is for the LLMs to produce hallucinations in order to analyse the construction of those, we opted for a simple prompt to translate the source and not to give any comments on the output (Sennrich et al., 2024).

Prompt: *Translate this text from English into Irish. Translate all words except named entities, and just respond with the translation, without any additional comments:* [full source text]

If the output contained untranslated words, we followed up with a secondary prompt to address the issue:

Follow-up Prompt: *The word(s) [untranslated word(s)] was/were not translated. Retranslate the full text making sure to translate these words. Just respond with the full text translation without any additional comments.*

Full texts were given so the model could make use of the whole context.[9]

4 Analysing Hallucinations in Irish

As noted previously, some characteristics of the Irish language, such as the heavy reliance on the use of suffixes and prefixes, and the great number of compound words, pose a great challenge for automatic translation. We observed a significant number of hallucinations related to verbs and nouns, adverbs and with fewer involving adjectives. Due to space and time constraints, we focus on hallucinations related to verbs and nouns. Table 2 presents the frequency of hallucinations across all test sets for both GPT4 and Mini.

The number of invented hallucinated words is greater for *nouns*, with only a few instances for *verbs*. The Mini model shows a greater number of invented hallucinated words in comparison with GPT4, showing a rate of 2.14 hallucinations of this type, against 0.86 hallucinations for the latter. This is an expected result regarding the model's performance, since Mini is a smaller and less robust version of GPT4. Smaller models generally have fewer parameters, which can impact their ability to accurately handle complex linguistic phenomena, such as Irish morphology and inflection. Nonetheless, since our objective is not to compare the models' outputs but rather to analyse the patterns in which these hallucinations are generated, differences in the number of hallucinations, as well as variations in model architecture and size, do not impact the validity of our study.

[9]Due to the Mini model's tendency to truncate longer texts, test set 3 was split into three segments, each retaining the opening to preserve key details (title, speaker, keywords).

Model	Verb	Noun	Total	Rate
GPT4	06	15	21	0.86
Mini	04	48	52	2.14
Total	10	63	73	-

Table 2: Frequency of invented-word hallucinations across test sets, with normalised rates expressed per 1,000 tokens.

Model	**Rules**	**No Rules**	**Total**	% Rules
GPT4	04	02	06	67
Mini	02	02	04	50

Table 3: Frequency of *VERB* hallucinations across all test sets. "Rules" denotes hallucinations conforming to Irish grammatical and morphological norms; "No Rules" denotes those that did not.

4.1 Hallucinating Irish Verbs

Table 3 presents the total number of hallucinated verbs and indicates whether they adhere to Irish grammatical and morphological rules. From the six invented hallucinated verbs by GPT4, four of them follow the Irish rules for grammar and morphology, and are classified as confabulations. Their application in context are shown in table 4.

We observe that when GPT4 confabulates verbs, its most common strategy is to reinterpret the source verb (e.g., 'sequenced', 'code', 'sequence') as a noun and then generate a corresponding Irish word. This results in the invention of forms such as *shraitheamar*, *códálann*, and *shraitheadh*, which, if they were actual Irish verbs, would be morphologically well-formed.

For example, in Example 1 in table 4, GPT4 has taken the noun *sraith*, meaning 'sequence' or 'series', and has correctly added the first person plural slender conjugation in the past tense, and lenitised the verb correctly, as is required in the past tense. In Example 2, GPT4 has taken the noun *cód* (which means 'code') and conjugated it using the correct broad present tense ending. However, it has added an additional syllable *ál*, which seems to align with the convention of verbs such as *tástáil* ('to test') which is conjugated as *tastálann* in the present tense. In Example 3, GPT4 has again taken the noun *sraith*, as in Example 1, and conjugated it into the past tense autonomous verb, the *briathar saor*. It has correctly lenitised the verb, as the direct relative clause particle '*a*' proceeds it.

Verbs	source	**output**	type
1	When we first sequenced this genome	Nuair a **shraitheamar** an géanóm seo ar dtús	conjugation of a noun
2	Triplets of those letters code for roughly 20	**Códálann** tripléid de na litreacha sin do thart ar 20	conjugation of a noun
3	so we could sequence them ..	go bhféadfaimis iad a **shraitheadh**	conjugation of a noun
4	Each device incorporating an Arm processor tends to be	**Tendeann** gach gléas a chuimsíonn próiseálaí Arm a bheith	English word conjugated

Table 4: Confabulated verbs that followed the Irish morphology rules by GPT4.o.

Verbs	source	**output**	type
1	it doesn't simulate the execution of code	nach **simulaíonn** sé comhoibriú cód	English word conjugated
2	Triplets of those letters code for roughly 20	**Códann** tríphéirí de na litreacha sin thart ar 20	Conjugation of a noun

Table 5: Confabulated verbs that followed the Irish morphology rules by GPT4.o. Mini

Model	**Rules**	**No Rules**	**Total**	%
GPT-4.0	11	04	15	73
Mini	19	29	48	39

Table 6: Frequency of *NOUN* hallucinations across all test sets. "Rules" denotes hallucinations conforming to Irish grammatical and morphological norms; "No Rules" denotes those that did not.

Example 4 shows another common type of confabulated verb. In this case, GPT4 adopts a well-documented feature of the Irish language — borrowing (Mulhall, 2018) words from English — while retaining the original English spelling and attempting to 'conjugate' them according to Irish grammatical patterns. *Tendeann* results from GPT4 taking the English verb 'tend' and correctly conjugates it into the first conjugation ending for slender vowels. There is no singular equivalent in Irish to the English verb 'tends to'.

While we decided that the listed examples are technically morphological, GPT4 also generated hallucinations that were not morphologically sound. For example, the verbal noun *athsraitheadh*. Here, the prefix *ath* (similar to 're-' in English) was applied to express the repetition of an action. However, while a lenition should typically follow a prefix in the stem of the verb, GPT4 omitted this. Another example of unnecessary omissions included the hallucination, *chog*. While seemingly attempting to translate the verb 'to chew', GPT4 omitted the latter half of '*chogain*' from its infinitive form and conjugated it into the first conjugation.

Regarding invented hallucinated verbs from the Mini model, from the four reported in table 2, two of them follow the Irish rules for grammar and morphology and are shown in table 5.

Similar to GPT4, the Mini model also generates confabulated verbs that follow two main patterns: transforming a source-language verb into a target-language noun, which then conjugated as if it were a verb, or retaining an English word while conjugating it according to Irish morphological rules. Example 1 in Table 5 illustrates the conjugation of the English verb 'simulate', by removing the third syllable and adding the correct present tense suffix *-aíonn*. Example 2 shows the conjugations of an Irish noun *cód* ('code') which has been used as the root of the verb and had a correct present tense ending of the first conjugation for broad vowels applied.

Invented hallucinated verbs which did not follow the rules were : *dearthach* which was the translation given for 'designing'. It appears that the model mistook 'designing' for an adjective and tried to translate it as that. The root *dear*, 'design' is correct, but in the second syllable it seems the model has combined the verbal adjective *deartha* and the suffix *-ach*, which commonly features in Irish adjectives. *Aknowimid* was the translation given for 'you know' (human translation: *tá a fhios agat/agaibh*) in the source text. *Aknowimid* uses the incorrect root, given that the Irish alphabet does not feature the letter 'k', and has been conjugated incorrectly using the slender first conjugation rather than the broad second conjugation. It seems that the model has attempted to say 'we acknowledge' even though it deviates slightly from the source to avoid a phrase that it was unfamiliar with.

4.2 Hallucinating Irish Nouns

As previously shown, the majority of hallucinated words in Irish were nouns. This is unsurprising given the intricacies of the five declensions of Irish nouns (see Appendix B). Table 6 presents the total number of hallucinated nouns generated by both models and indicates whether they adhere to Irish grammatical and morphological rules. To better structure the analysis of these hallucinated nouns, this section is divided into the following types: Compounds (section 4.2.1), Lazy Gaelicisation (section 4.2.2), Good Hallucination (section 4.2.3), Code-switching (section 4.2.4), Prefix (section 4.2.5), and Suffix(section 4.2.6).

	source	GPT4.o
1	...results of independent performance benchmarks...	...torthaí de **bhinncheisteanna** feidhmíochta neamhspleácha...

Table 7: Confabulated **Compound Nouns** that followed the Irish morphology rules by GPT4.o

	source	Mini
1	Or, in this case, windmill.	Nó, sa chás seo, **gaothmhoill**.
2	Evolution of the turbine	Evoláid na **gaothchumhachta**
3	...modern wind turbines are huge...	...tá **gaothmhoillí** nua-aimseartha ollmhóra...
4	Wind turbines are reaching ever higher.	Tá **gaothchumhachtaí** ag dul níos airde agus níos airde.
5	results of independent performance benchmarks	torthaí na **gcomhairlíon** próiseálaí neamhspleácha

Table 8: Confabulated **Compound Nouns** that followed the Irish morphology rules by GPT4.o Mini.

	source	GPT4.o
1	...on all of the elements in the nacelle.	...ceann de na heilimintí sa **nascáil**.
2	Triplets of those letters code for roughly	Códálann **tripléid** de na litreacha sin do thart ar
3	...what we're calling combinatorial genomics	...atá á ghlaoch againn **géanómóireacht** chomhcheangailteach

Table 9: Confabulated words that followed the Irish morphology rules by the GPT4.o classified as '**Lazy Gaelicisation**".

4.2.1 Compounds

Both models have used a compounding of nouns to create invented words. Table 7 illustrates the one instance of compounding of two nouns in GPT4 *bhinncheisteanna* which compounds the noun *binn* ('peak', 'cliff' or 'edge') and *ceisteanna* ('questions'). In Irish, compounding often involves initial consonant mutations in the second or subsequent parts of the compound (Ball and Muller, 2010, 176). Therefore, the hallucinated word *bhinncheisteanna* follows this pattern correctly, applying lenition to the second component, *c***h***eisteanna*. No other compound nouns, either morphologically correct or incorrect, was invented by this model.

Regarding invented compounds by the Mini model, table 8 illustrates the 5 confabulated examples that could be classified as morphologically correct, although they carry little meaning. Example 1 *gaothmhoill* (attempted translation of 'windmill') is a compounding of the word *gaoth* ('wind') and *moill* ('delay'). The morphological rule of initial consonant mutations (lenitions) is followed. A pattern emerged in the hallucinations created for this category in Mini, whereby, the first noun in the compound is correct or relates to the source text, but is followed by an incorrectly translated noun. The second noun *moill* is nonsensical in this context, however, it does resemble the English noun 'mill'. Example 2 *gaothchumhachta* (attempted translation of 'turbine and wind turbine') compounds *gaoth* ('wind') and *cumhacht* ('power'). This translation differs greatly from the human translation *tuirbín* and *tuirbín gaoithe*. A lenition is applied correctly to the second noun and it is correctly in the genitive singular in all 9 cases, as is required. Example 3 *gaothmhoillí* is similar to example 1, but the second noun is in the plural. However, the word to be translated in Example 1 is 'windmill', in contrast to 'wind turbines' in Example 3. Example 4 is similar to Example 2, as it also compounds *gaoth* and *cumhacht*, however, the second noun is in the nominative and genitive plural, which is correct in all 6 cases. Example 5 shows *comhairlíon* as a translation for 'benchmarks'. It compounds *comhair* ('combined work', 'co-operation', 'partnership'), with *líon* (a full number, complement). It is morphologically correct, as it follows orthographic rules (broad vowels followed by broad vowels, slender vowels followed by slender). A lenition is not added to the second word, as a lenition cannot be added to an 'l'. Example 5 shows less logic than the other pattern and seems to compound two random nouns to create an invented hallucination.

The only invented hallucination for compounded nouns by the Mini model that did not follow morphological rules was *gaoithchumachta*, which while similar Example 2 in Table 8 contains and 'i' in *gaoth*, meaning it does not follow orthographic rules.

4.2.2 Lazy Gaelicisation

We refer to instances where translations appear to have been generated based on the phonetics of the English word, often modifying the spelling to conform to Irish orthographic rules even though a corresponding word exists in Irish as *Lazy Gaelici-*

	source	Mini
1	It will handle turbine blades...	Rachaidh sé i ngleic le **blaide** gaothchumhachta...
2	so we thought we'd build them in cassettes...	mar sin shocraíomar iad a thógáil i **gcásáidí**...
3	...this may sound like genomic alchemy...	...b'fhéidir go mbeidh sé seo cosúil le **alcaimíocht** ghéineamach...
4	Now I've argued, this is not genesis;	Anois, rinne mé argóint, ní **ghinéise** atá anseo;

Table 10: Confabulated words that followed the Irish morphology rules by the GPT4.o Mini classified as '**Lazy Gaelicisation**".

	source	GPT4
1	...heart of a device controller, a microcontroller (MCU)	...chroílár rialtóra gléas, **micriríalt óir** (MCU)
2	this is just a regular photomicrograph.	níl anseo ach **fótamhicreagraf** rialta.
3	...with synthetic bacteria, Archaea...	...le baictéir shintéiseacha, **Seanríochtaí**...

Table 11: '**Good**' Confabulated words that followed the Irish morphology rules by GPT4.o

	source	Mini
1	Giant fans of wind energy	**Fanaithe** ollmhóra de fuinneamh gaoth
2	...in what sources outside of Apple call an "emulator"	...ar a dtugtar "**simulachtóir**"...
3	...invention, science, technology	...**inventiú**, eolaíocht, teicneolaíocht

Table 12: Confabulated words with **code-switching**, that is, English nouns that followed the Irish morphology rules for by GPT4.o Mini.

sation. Both engines (GPT4 in table 9, and the Mini model in table 10) produced confabulations of this variety. Many of these confabulated words could plausibly be mistaken for legitimate Irish terms, particularly in casual reading. At the very least, the reader would recognise their connection to the English source and infer the intended meaning with relative ease. These phonetic adaptations have been found among Irish speakers, particularly in informal or spontaneous speech, and sometimes in writing (Darcy, 2014). The GPT4 model presented a few of those cases. These examples represent a clear alignment with our definition of confabulation: invented words that are not simply erroneous, but exhibit internal coherence and plausibility according to Irish norms.

In Example 1 in table 9, the model translated the word 'nacelle' as *nascáil* ('linkage') instead of *naoisil*, which is the correct translation. We note that both terms ('nacelle' and *nascáil*) are extremely phonetically similar which could explain this hallucination. Moreover, 'nacelle' is highly specialised language relating to aeronautical engineering, and therefore, it is entirely possible that the Irish term is newly coined, after the model was last updated as the national terminology database for Irish is updated constantly.[10] In Example 2, GPT4 translated 'triplets' as *tripléid* (correct translation is *tríríní*), in which the second syllable *-pléid* demonstrates a correct pluralisation. 'Triplet' and *tripléid*, which we assume the model believes is the singular, are phonetically similar justifying the model's reasoning. In example 3, GPT4 translated 'genomics' as *géanómóireacht* (correct translation is *géanómaíocht*). The model has added an unnecessary syllable, however the reasoning is unclear.

The Mini model also presented a few examples of *Lazy Gaelicisation* as shown in table 10, translating 'blades' as *blaide* (correct translation is *lanna*), following the orthographical rules by matching the slender vowels. In example 2, the model translated 'cassettes' as *cásáidí* (correct translation is *caiséid*). Irish nouns ending in *-áid* are usually feminine, belonging to the 2nd declension. Therefore, they are pluralised using the suffix *-í*, as the model has done. The Irish noun *caiséad* is masculine in the 1st declension, meaning the last consonant must be slenderised to produce the plural (both nominative and genitive case). Example 3 shows the translation of the word 'alchemy' as *alcaimíocht* (correct translation is *ailceimice*). The Irish suffix *-(a)íocht* is commonly used to express the English suffix '-ation'. For example, *reachtaíocht*, 'legislation', *eagraíocht*, 'organisation', *radaíocht*, 'radiation', *cúrsaíocht*, 'circulation'(of money), *cáilíocht*, 'qualification'. It is possible that the model took the context of the test set into consideration and was influenced by domain-specific nouns that it was familiar with. In example 4, the model translated the term 'genesis' as *ginéise*. The intended meaning in the source text refers to the beginning of something,

[10] https://www.tearma.ie/

therefore *bunús* is used to express this in Irish.

Examples in this category from the Mini model that do not follow morphological rules include: *protáitíopaíocht* (human translation *fréamhshamhaltú*) and *protáitíopaí* (human translation *fréamhshamhlacha*), used to translate the terms 'prototyping' and 'prototypes' respectively, used the incorrect prefix for 'proto-'; *evólúisian* as the translation for 'evolution', disregards the convention of the Irish alphabet which does not include the letter 'v'; *autagrafaí*, used to translate 'autograph', uses the incorrect prefix for 'auto-' which is usually *uath-*. In this case, the correct translation is *átagraf*.

4.2.3 Good Confabulations

We classify good confabulations as invented outputs that seem to follow all morphological rules for words which had no official translation available, but a good attempt has been made to create a word. These cases demonstrate creative yet coherent language generation in the absence of concrete lexical data. There were 3 such cases produced by GPT4 (see table 11), and no cases by the Mini model.

Example 1 shows *micririaltóir* as a translation of 'microcontroller' and correctly compounds the prefix *micri-* with the noun *rialtóir* (person). While this is a good attempt, the correct 'controller' in this context would be *rialtán* ('switch', 'button', 'dial'). There is no lenition added following the prefix, as lenitions cannot be added to the letter 'r'. Example 2, a translation of 'photomicrograph' shows a similar pattern to Example 1. The prefix *fóta-* is correctly added to the noun *micreagraf*, and a lenition is correctly added following the prefix. Example 3, *Seanríochtaí* as a translation for Archea is interesting, as it compounds the adjective *sean* ('old') with the noun *ríochtaí* ('kingdoms'). This is of interest as it appears to use an understanding of *Archea* as the adjective 'archaic' and translates it as such to *sean*.

Deceiving 'good hallucinations'- are invented hallucinations which, similar to *Lazy Gaelicisation*, seem and sound like correct Irish words, but upon further inspection, carry no meaning. This is the case of the word *laigeas* (produced by the Mini model), in an attempt to translate 'bending moments' from the source text while *Laigeas* appears to be a morphologically correct word, it contains no real units of meaning.[11]

[11] source: 'able to withstand bending moments up to 100.000 kNm", output: '...atá in ann **laigeas** a fhulaingt suas le 100.000 kNm'.

4.2.4 Code-switching

We look into examples where the models have taken an English noun and added an Irish suffix in an attempt to create an Irish word. This phenomena has been reported in the use of Irish in tweets and been classified as code-switching word-level alternation (Lynn and Scannell, 2019). These examples differ from Lazy Gaelicisation in that they appear to be a compounding between the source language and the target language, disregarding the orthographical conventions of the Irish language. They also illustrate another facet of confabulation, where the system fills lexical gaps by improvising plausible word forms, albeit in ways that stretch or break conventional language norms. There were no occurrences of English nouns with Irish suffixes in GPT4.

In table 12 Example 1 *fanaithe*, the Mini model has taken the English noun 'fan' and added the Irish suffix *-aithe* which is commonly used to pluralise broad weak plural Irish nouns. In Example 2, *simulachtóir*, the model took the first two syllables of the noun 'simulator' and added the Irish suffix used to express '-ator', *-achtóir*. Regardless this is a mistranslation as the source calls for 'emulator'. In Example 3, an attempt to translate the word 'invention' into *inventiú* was made by taking the first two syllables of the noun and adding the Irish suffix *-iú*.

One xample that do not follow Irish rules was *Simuláid*, which was an attempted translation of 'simulation'. In this case, the root 'simul' is not morphologically acceptable, the suffix *-áid* is seen across Irish in other nouns such as *cumarsáid* ('conversa**tion**') and *oráid* ('ora**tion**').

4.2.5 Prefix

Table 13 shows examples of hallucinations in which the Mini model created nouns using the correct prefixes established within Irish morphology. In Examples 2 and 3, the model appeared to recognise the prefixes in their source form and translated them to Irish without correctly translating the latter parts of the nouns. Example 1 shows an attempt to have a similar function of the meaning of the source noun. GPT4 model did not hallucinate any words with a 'correct' prefix.

Both GPT4 and the Mini model confabulated nouns with prefixes that were phonetically similar but incorrectly spelt. For example, *micoplásma* in-

	source	Mini
1	mainly in the area of composites.	go príomha i réimse na **gcomhshamhlacha**.
2	Triplets of those letters code for roughly 20 amino acids,	Códann **tríphéirí** de na litreacha sin thart ar 20 aigéad aimín,
3	we think that biology can have a major impact	gur féidir leis an **bithleacht** níos mó tionchar a imirt

Table 13: Confabulated words that followed the Irish morphology rules for **prefix** by GPT4.o Mini.

	source	GPT4
1	...there's a problem when it comes to simulating wind turbines.	...tá fadhb ann maidir le **turasáin** gaoithe a insamhladh.
2	...forces and moments on the shaft in three directions.	...fórsaí agus **cuimhneachtaí** ar an seafta i dtrí threoir.
3	that can take three million rads of radiation.	is féidir a ghlacadh trí mhilliún **radaim** radaíochta.
4	Archaea and, eventually, eukaryotes.	Seanríochtaí agus, faoi dheireadh, **eocaróitigh**.

Table 14: Confabulated words that followed the Irish morphology rules for **suffix** by GPT4.o.

	source	Mini
1	starting with the digital information of the genome of phi X174.	ag tosú leis an eolas digiteach de **ghéineomaí** phi X174.
2	that can take three millions rads of radiation.	atá in ann trí mhilliún **radán** de radaíocht a ghlacadh.
3	we can select for viability...	is féidir linn roghnú le haghaidh **feidhmeannaíochta**...

Table 15: Confabulated words that followed the Irish morphology rules for **suffix** by GPT4.o Mini.

stead of *míceaplasma*, *cilivata* instead of *cileavata*. In other instances, both systems created hallucinations by keeping the prefix in its source form and translating the rest of the noun. For example, *megavata* and *sub-aonadanna*.

4.2.6 Suffix

The following hallucinations were identified and characterised by their use of real Irish nouns and the addition of an infix or suffix for a certain purpose.

All listed hallucinations generated by GPT4 are concerned with pluralised nouns (table 14). Examples 1 and 3 show the inclusion of an infix in order to pluralise nouns, while the hallucinations that occurred in Examples 2 and 4 applied a suffix. These confabulations are deemed morphologically correct as they are typical of Irish spelling and also respect the conversions set out in the declensions.

The Mini model (table 15) generated confabulations that show phonological similarities to their correct translation, however the addition of suffixes could only be deemed unnecessary. In Example 1, the model produced the suffix *-aí* in *ghéineomaí*, which is commonly used to indicate a particular person or job in Irish. A possible explanation for this is that the Mini model may have misunderstood the source 'genome' to be an agent, rather than an object. Examples 2 and 3 show hallucinations where the first parts of the noun are correct, however noun endings that are common within Irish morphology were added. Interestingly, despite their incorrect endings, these nouns still respect the grammatical rules that are involved when counting items and turning a noun into its genitive case form.

Both systems also generated hallucinations where an apparent disconnect occurred between their spelling and patterns of mutations. For example, while attempting to translate 'voltage dips', GPT4 generated *dippaí*. This was deemed morphologically incorrect as it took the source noun, which is an existing loan word in Irish, that does not differ in the singular for the English 'dip', however, the correct plural is *dipeanna*. In this case GPT4 added double consonants (pp) and a strong plural ending. The Mini model created incoherent hallucinations such as *dhearadhóir* in an attempt to translate 'designer'. The model took the Irish *dearadh*, meaning 'design' and attached a suffix that offers the same function as '-er'(*-óir*) in English to suggest an agent. This hallucination, however, was not deemed morphologically correct as it does not align with spelling conventions.

5 Discussion and Conclusions

This study examined the types of hallucinations involving the creation of new words in LLM-generated translations into Irish and evaluated whether these hallucinations adhered to Irish linguistic rules, and therefore classified as confabulations. Our findings indicate that both GPT4 and Mini exhibit similar patterns of word invention, though the latter produces hallucinated words at a significantly higher frequency. While both models demonstrate a tendency to confabulate, that is, apply Irish morphological rules to these hallucinated words, GPT4 adheres to these rules more consistently (71%) than the Mini model (40%) (Tables 3 and 6). This difference likely reflects the Mini model's smaller size and reduced robustness. Nonetheless, both models produce *plausible* but *non-existent* lexical items that raise intriguing ques-

tions about their potential influence of confabulations on the Irish language.

Many of the confabulations resemble patterns made by learners of Irish, such as code switching and borrowings. This suggests that the models might not be generating entirely arbitrary forms but are instead applying Irish word formation rules in a way that mirrors natural language learning processes. These confabulation patterns are particularly relevant in the context of what Fhlannchadha and Hickey (2018, p.21) describe as a 'post-traditional variety of Irish'—a variety adopted by non-native speakers who do not align with any particular dialect of Irish. The authors note that established ideologies rooted in native and traditional models of Irish are being disrupted by new speakers, creating a notable tension between linguistic groups in the era of language revitalisation. They caution that the expansion of post-traditional Irish could lead to the erosion of crucial aspects of the language, particularly in lexicon and grammar. Similar concerns arise in other morphologically rich, low-resource languages, such as Scottish Gaelic, and Welsh, where language change and revitalisation efforts interact with evolving speaker communities. In this light, LLM-generated confabulations raise further questions about the role of AI in reinforcing or reshaping these dynamics across such languages.

But what does it mean when an AI model exhibits patterns akin to human learners? Could these errors, if encountered frequently in machine-generated content, influence the way Irish is written or even spoken over time? Two particularly noteworthy categories of confabulations observed in this study, which we term 'Lazy Gaelicisation' and 'Good Confabulation', involve the adaptation of English words into Irish-like phonetics, often by modifying their spelling to align with Irish orthographic rules. This phenomenon is not exclusive to LLMs; similar strategies have been observed among Irish speakers themselves. The phonetic adaptation of English words into Irish structures has long been a feature of the language, seen both in historical borrowings and in contemporary informal speech. Does this suggest that such hallucinations are merely an extension of a natural linguistic process? Or should they be viewed as problematic, reinforcing patterns of language shift rather than supporting authentic Irish usage?

Models invented words that follow morphological rules when no official translation is available ('good confabulations'). The introduction of novel, non-standard words could be seen as either a sign of language erosion or a potential source of linguistic innovation. While some have found the replacement of existing Irish words with English-derived forms as a form of 'detrimental change' (Hickey, 2009, 671), others see partial or non-standard Irish as a step toward broader engagement. If LLM-generated forms gain traction, could they help fill lexical gaps in technical domains where Irish terminology is scarce? Or would they risk further undermining existing Irish vocabulary? These are not straightforward questions, and rather than offering definitive answers, they highlight the need for continued observation and discussion.

Finally, it is important to highlight the specific context in which these hallucinations and confabulations occur. In our study, most invented hallucinated words appeared in technical and specialised domains, where even fluent speakers may struggle with terminology. In more general texts, where Irish has a more established lexicon, the models produced fewer invented words, although overall grammatical accuracy and fluency remained an issue. This suggests that while hallucinations in LLM-generated translations may be concerning in certain contexts, their broader impact on Irish will likely depend on how these models are used and integrated into real-world workflows. Future research should explore these issues further with larger datasets and more extensive replication resources, particularly regarding the long-term implications of LLM-assisted translation for minority languages. Moreover, future work should look into how speakers perceive and react to these hallucinations and confabulations.

Nonetheless, this work serves as a foundation for further investigations into the implications of LLM errors in morphologically rich, low-resource languages. Our goal is to encourage discussion on the long-term impact of these technologies on language, especially in the case of low-resource, morphologically rich languages.

Acknowledgments

We thank Prof. Ciarán Mac Murchaidh for the invaluable help and discussions. The first author benefits from being member 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

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.

Martin Ball and Nicole Muller. 2010. *The Celtic Languages*, 2nd edition. Taylor Francis, Hoboken.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Eleftheria Briakou, Zhongtao Liu, Colin Cherry, and Markus Freitag. 2024. On the implications of verbose llm outputs: A case study in translation evaluation. *Preprint*, arXiv:2410.00863.

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. *arXiv preprint*. Issue: arXiv:2005.14165 arXiv:2005.14165 [cs].

Lauren Cassidy. 2024. *Linguistic analysis and automatic dependency parsing of Tweets in modern Irish*. Phd thesis, Dublin City University.

Sheila Castilho, João Lucas Cavalheiro Camargo, Miguel Menezes, and Andy Way. 2021. DELA Corpus - A Document-Level Corpus Annotated with Context-Related Issues. In *Proceedings of the Sixth Conference on Machine Translation*, pages 571–582. Association for Computational Linguistics (ACL).

Sheila Castilho, Clodagh Quinn Mallon, Rahel Meister, and Shengya Yue. 2023. Do online machine translation systems care for context? what about a GPT model? In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 393–417, Tampere, Finland. European Association for Machine Translation.

Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? 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 536–541, New Orleans, Louisiana. Association for Computational Linguistics.

Guinevere Darcy. 2014. *Code-mixing and context: A Corca Dhuibhne case study*. Unpublished phd dissertation, University of Limerick.

Meghan Dowling, Teresa Lynn, Alberto Poncelas, and Andy Way. 2018. Smt versus nmt: Preliminary comparisons for irish. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 12–20.

Meghan Dowling, Joss Moorkens, Andy Way, Sheila Castilho, and Teresa Lynn. 2020. A human evaluation of english-irish statistical and neural machine translation. In *22nd Annual Conference of the European Association for Machine Translation*, page 431.

Siobhán Nic Fhlannchadha and Tina M. Hickey. 2018. Minority language ownership and authority: perspectives of native speakers and new speakers. *International Journal of Bilingual Education and Bilingualism*, 21(1):38–53.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

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*.

Tina Hickey. 2009. Code-switching and borrowing in irish. *Journal of Sociolinguistics*, 13(5):670–688.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.* Just Accepted.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

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.

Séamus Lankford, Haithem Afli, and Andy Way. 2023. adaptmllm: Fine-tuning multilingual language models on low-resource languages with integrated llm playgrounds. *Information*, 14(12).

Séamus 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.

Teresa Lynn and Kevin Scannell. 2019. Code-switching in Irish tweets: A preliminary analysis. In *Proceedings of the Celtic Language Technology Workshop*, pages 32–40, Dublin, Ireland. European Association for Machine Translation.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

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.

Chris Mulhall. 2018. Irish lexicography in borrowed time: The recording of anglo-irish borrowings in early twentieth-century irish dictionaries (1904-1927). *International Journal of Lexicography*, 31(2):214–228.

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.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 21–33, St. Julian's, Malta. Association for Computational Linguistics.

Peng Shu, Junhao Chen, Zhengliang Liu, Hui Wang, Zihao Wu, Tianyang Zhong, Yiwei Li, Huaqin Zhao, Hanqi Jiang, Yi Pan, Yifan Zhou, Constance Owl, Xiaoming Zhai, Ninghao Liu, Claudio Saunt, and Tianming Liu. 2024. Transcending language boundaries: Harnessing llms for low-resource language translation. *Preprint*, arXiv:2411.11295.

Peiqi Sui, Eamon Duede, Sophie Wu, and Richard So. 2024. Confabulation: The surprising value of large language model hallucinations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14274–14284, Bangkok, Thailand. Association for Computational Linguistics.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang Nguyen. 2024a. Irish-based large language model with extreme low-resource settings in machine translation. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 193–202, Bangkok, Thailand. Association for Computational Linguistics.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D Nguyen. 2024b. Uccix: Irish-excellence large language model. *arXiv preprint arXiv:2405.13010*.

A Appendix A - Irish Verbs

B Appendix B - Irish Declensions

1st Conjugation	Broad vowels	Slender vowels	**2nd Conjugation**	Broad vowels	Slender vowels
1st sing	*-aim*	*-im*	1st sing	*-aím*	*-ím*
2nd sing	*-ann tú*	*-eann tú*	2nd sing	*-aíonn tú*	*-íonn tú*
3rd sing	*-ann sé/sí*	*-eann sé/sí*	3rd sing	*-aíonn sé/sí*	*-íonn sé/sí*
1pl	*-aimid*	*-imid*	1pl	*-aímid*	*-ímid*
2nd pl	*-ann siad*	*-eann sibh*	2nd pl	*-aíonn sibh*	*-íonn sibh*
3rd pl	*-ann siad*	*-eann siad*	3rd pl	*-aíonn siad*	*-íonn siad*
Aut.	*-tar*	*-tear*	Aut.	*-aítear*	*-ítear*

Table 16: Suffixes for Conjugation of Irish Verbs in the Present Tense

Declension	Gender	**Nominative Singular**	**Genitive Singular**
1st	*M*	*Ends on a broad consonant*	Last consonant is slenderised
2nd	*F (except for im, sliabh)*	*Ends on a consonant either broad or slender*	Ends with '-e'
3rd	*M & F*	*Ends on a consonant either broad or slender*	Ends with '-a'
4th	*M & F*	*Ends with a vowel or '-ín'*	Remains the same as the nominative singular
5th	*F (few M)*	*Ends with '-il', '-in', '-ir' or a vowel*	Ends on a broad consonant
Declension	Gender	**Nominative Plural**	**Genitive Plural**
1st	*M*	*Same form as the genitive singular*	*Same form as the nominative singular*
2nd	*F (except for im, sliabh)*	*Ends with '-a', e.g. bróga, scornacha*	*Loses the '-a', e.g. bróg, scornach*
3rd	*M & F*	*Ends with '-a', '-acha', '-(a)í', '-(e)anna', '-ta', '-te'*	
4th	*M & F*	*Ends with '-(a)í', '-(e)anna', '-(i)te', '-(i)the', '-nna'*	
5th	*F (few M)*	*Ends with '-(e)acha', '-idí', '-na', '-ne'*	

Table 17: Verb Declension

Patent Claim Translation via Continual Pre-training of Large Language Models with Parallel Data

Haruto Azami[1] **Minato Kondo**[1] **Takehito Utsuro**[1] **Masaaki Nagata**[2]
[1]Deg. Prog. Sys.&Inf. Eng., Grad. Sch. Sci.&Tech., University of Tsukuba
[2]NTT Communication Science Laboratories, NTT Corporation, Japan
s2420710_@_u.tsukuba.ac.jp, s2320743_@_u.tsukuba.ac.jp,
utsuro_@_iit.tsukuba.ac.jp,
masaaki.nagata_@_ntt.com

Abstract

Recent advancements in large language models (LLMs) have enabled their application across various domains. However, in the field of patent translation, Transformer encoder-decoder based models remain the standard approach, and the potential of LLMs for translation tasks has not been thoroughly explored. In this study, we conducted patent claim translation using an LLM fine-tuned with parallel data through continual pre-training and supervised fine-tuning. A comparative evaluation against Transformer encoder-decoder based translations showed that the fine-tuned LLM achieved high scores for both BLEU and COMET, demonstrating improvements in addressing issues such as omissions and repetitions. Nonetheless, hallucination errors, which were not observed in traditional models, occurred in some cases and negatively affected translation quality. These findings highlight the promise of LLMs for patent translation while also identifying challenges that warrant further investigation.

1 Introduction

Large language models (LLMs) demonstrate exceptional versatility because of their extensive pre-training, proving highly effective in various natural language processing tasks, such as summarization and question-answering. In the field of machine translation, closed LLMs like GPT-4 have been reported to achieve higher human evaluation scores than existing translation models (Kocmi et al., 2023, 2024). However, in the patent domain, Transformer encoder-decoder based translation models remain the mainstream approach, and the translation capabilities of LLMs have not been sufficiently explored. The translation quality of patent documents has reached a sufficiently high level with conventional neural machine translation (NMT) methods, particularly for the main body of patent texts. However, patent claims remain a notable exception where translation quality is still problematic. Patent claims are known for their extremely long and syntactically complex sentence structures, which pose significant challenges for traditional models. In addition, this study focuses on Japanese-to-English translation, where a major obstacle is the significant difference in word order between the two languages. Such structural divergence further complicates the translation of patent claims, especially in preserving the meaning and consistency across long sequences. In contrast, LLMs are believed to be capable of translating long sequences while maintaining global coherence and consistency. Motivated by this potential, the present study investigates how effectively LLMs can translate patent claims, which represent the most difficult component in patent translation. To this end, we adopt the method proposed by Kondo et al. (2024), utilizing parallel patent data for continual pre-training and supervised fine-tuning (SFT) to construct an LLM specialized in patent claim translation. The performance of this LLM is then compared with that of conventional Transformer-based models, with translation quality evaluated using metrics such as BLEU and COMET. The results demonstrate that the LLM statistically significantly outperforms conventional models, effectively addressing issues such as omissions, repetitions, and terminology inconsistency. However, the study also reveals LLM-specific challenges, such as hallucinations, which are observed in specific cases that do not occur in conventional models. This study evaluates both the potentials and challenges of applying LLMs to patent translation, highlighting their effectiveness and identifying areas requiring further improvement.

Proceedings of Machine Translation Summit XX Volume 1, pages 300–314, June 23-27, 2025

2 Related Work

2.1 Translation of Patent Claims

Patent claims are one of the most important parts of a patent document, and they are characterized by strict sentence structures and specialized terminology, making them a significant challenge for machine translation.

Fuji et al. (2015) applied statistical machine translation (SMT) to the translation of English, Chinese, and Japanese patent claims and proposed a method for appropriately transforming claim structures. Their approach utilized manually created synchronous context-free grammar (SCFG) rules to convert the source language structure into the target language structure, thereby addressing the unique descriptive style found in patent claims. However, this method had a limitation: the need for manual rule creation that hindered the flexible adaptation to new descriptive styles.

Additionally, research on patent claim translation has been explored in the NTCIR patent translation task. Conducted by Fujii et al. and Goto et al. from 2008 to 2013, respectively, this task primarily employed SMT, advancing the use of parallel corpora and evaluation methods for patent document translation. In particular, translating lengthy patent claims requires maintaining consistent terminology and proper structural transformations, often supplemented by rule-based approaches.

Subsequently, the patent translation task was incorporated into the Workshop on Asian Translation (WAT), where the neural machine translation (NMT) approach, which had already become dominant in machine translation, was applied to patent translation, as demonstrated by Nakazawa et al. (2016). While NMT improved translation fluency, maintaining the strict structure of patent claims remained a challenge. In recent years, there has been progress in constructing large-scale parallel corpora specifically for patent translation. In 2022, the EuroPat corpus was released by K. Heafield and Wiggins (2022), providing a multilingual parallel dataset based on European patent documents. This resource laid a foundation for research in patent translation, especially among European languages. More recently, in 2024, JaParaPat—a large-scale Japanese-English parallel corpus for patent translation—was introduced (Nagata et al., 2024). Constructed using patent family alignments between Japanese and U.S. patent applications, this resource is utilized in our study as training data for both the continual pretraining and supervised fine-tuning of LLM. The development of such domain-specific resources facilitates research aimed at improving patent translation quality, particularly for the Japanese-English language pair.

2.2 LLM-based Translation

In recent years, LLMs have gained attention in the field of machine translation, demonstrating high accuracy in general text domains such as news articles and dialogues. In particular, the use of QLoRA for fine-tuning LLMs has significantly improved multilingual translation performance (Zhang et al., 2023).

Guo et al. (2024) and Kondo et al. (2024) proposed a method combining continual pre-training on parallel data with SFT to enhance the LLM-based translation performance beyond the traditional Transformer encoder-decoder based models. Their approach involved the continual pretraining using large-scale web-crawled parallel corpora, followed by SFT with high-quality parallel datasets, notably improving translation accuracy. Specifically, Kondo et al. (2024) provided a detailed analysis of the Japanese-English translation, addressing the dataset selection and fine-tuning strategies.

In parallel, recent work has explored domain adaptation methods tailored for LLM-based machine translation. Zheng et al. (2024) conducted a comprehensive comparison of fine-tuning strategies such as full fine-tuning, LoRA, and prompt tuning, demonstrating their effectiveness in adapting LLMs to domain-specific translation tasks. Moslem et al. (2023) proposed an adaptive machine translation framework using LLMs, which integrates context-aware prompting and auxiliary data to improve translation quality in specialized domains. These studies highlight the growing interest in leveraging LLMs for translation in complex, domain-specific settings such as legal or patent language, which motivates our focus on patent claim translation using domain-adapted LLMs.

Recent research also points out key challenges and refinements in LLM-based translation. Xu et al. (2024) demonstrated that models predominantly pre-trained on English data, such as LLaMA-2, suffer reduced translation accuracy when translating into non-English target languages. To address this, they introduced ALMA,

a two-stage fine-tuning method: first with monolingual data, then with a small quantity of high-quality parallel data.

Despite these advances, LLM-based translation models have primarily been evaluated on test sets from the WMT General Machine Translation Task (Kocmi et al., 2022, 2023) and Flores-200 (Team et al., 2022), and their effectiveness across diverse domains remains underexplored.

3 Experimental Setup

3.1 Model and Training Procedure

This study follows the approach of Kondo et al. (2024), applying continual pre-training and supervised fine-tuning (SFT) to an open-source LLM, **rinna/llama-3-youko-8b**[1], hereafter referred to as **youko-8b**. youko-8b is a 7B-parameter model initially pre-trained on 22 billion tokens of Japanese and English monolingual data. To adapt the model to the patent translation task, we conducted continual pre-training using parallel patent data, followed by SFT to specialize it for translating patent claims.

3.2 Dataset

We used JaParaPat (Nagata et al., 2024), a large-scale Japanese-English parallel corpus of patent data, for both continual pre-training and supervised fine-tuning. JaParaPat consists of approximately 300 million sentence pairs constructed from patent applications published between 2000 and 2021 by the Japan Patent Office (JPO) and the United States Patent and Trademark Office (USPTO). The dataset was created through document alignment based on patent family information, followed by sentence segmentation and machine-translation-based sentence alignment.

In this study, different subsets of JaParaPat were used depending on the purpose:

- For continual pre-training, we used parallel data from 2016 to 2020, comprising approximately 61 million sentence pairs. From this data, 50,000 sentence pairs were excluded to construct a development set. Sentence similarity was calculated using LaBSE (Feng et al., 2022), and 10,984 pairs with similarity scores between 0.9 and 0.95 were selected as the development set.

[1] https://huggingface.co/rinna/llama-3-youko-8b

Usage	Time Period	Data Type	Sentence Pairs	English Words
continual pre-training	2016∼2020	training	61,364,685	1.9B
		development	10,984	327K
SFT	2021	training	15,000	53.6K
		development	1,000	36.7K
test set	2021	—	33,923	—

Table 1: Usage and Details of Patent Parallel Data

- For supervised fine-tuning, we used the 2021 portion of JaParaPat, focusing on patent claims. Sentence pairs were filtered based on similarity scores (0.8 to 0.95), and the selected data was divided into training and development sets. The test set was also constructed from 2021 patent claims by selecting unique sentence pairs with similarity scores between 0.9 and 0.95 and containing more than 100 words.

Table 1 summarizes the breakdown of the data used in each stage.

The input format for continual pre-training was as follows:

```
{Japanese sentence}
{English sentence}
```

For supervised fine-tuning, we used a prompt-based format:

> これを日本語から英語に翻訳してください.
>
> 日本語 (Japanese):Japanese sentence
>
> 英語 (English):English sentence

The English translation of the above prompt is:

> "Translate this from Japanese to English."

We applied both full fine-tuning and LoRA (Hu et al., 2022) for the supervised fine-tuning stage.

3.3 Hyperparameter Settings

The hyperparameters of the continual pre-training are shown in Table 2, and the hyperparameters of the SFT are shown in Table 3. In continual pre-training, bfloat16 and DeepSpeed ZeRO stage 2 (Rasley et al., 2023) were applied during training. The SFT was performed on the model that achieved the lowest validation error during the continual pre-training.

Hyperparameter	Value
optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$)
learning rate schedule	cosine scheduler
warmup ratio	1%
max learning rate	2.5×10^{-5}
weight decay	0.1
gradient Clip	1.0
batch Size	1,024
validate interval updates ratio	10%
epochs	1

Table 2: Hyperparameters for Continual Pre-training

Hyperparameter	Value
optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$)
learning rate schedule	cosine scheduler
warmup ratio	1%
max learning rate	2.5×10^{-6}
weight decay	0.1
gradient Clip	1.0
batch Size	64
epochs	2

Table 3: Hyperparameters for Supervised Fine-Tuning

3.4 Comparative Methods

3.4.1 Baseline

As a baseline, we employed a Transformer encoder-decoder based translation model. The model was trained on the same patent parallel corpus as the LLM-based models, comprising approximately 61M sentence pairs. Specifically, we employed the machine translation software by Fairseq (Ott et al., 2019) and used Transformer Big (Vaswani et al., 2017) as the translation model. The hyperparameters of the Transformer model are shown in Table 4. The training and test data were tokenized using SentencePiece (Kudo and Richardson, 2018), which was trained on a random sample of 10M sentence pairs from the patent parallel corpus. The vocabulary size was set to 32K for both Japanese and English.

Hyperparameter	Value
architecture	Transformer_vaswani_wmt_en_de_big
enc-dec layers	6
optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$)
learning rate schedule	Inverse square root decay
warmup steps	4,000
max learning rate	0.001
dropout	0.3
gradient Clip	1.0
batch Size	16K tokens
max number of updates	60K steps
validate interval updates	1K steps

Table 4: Hyperparameters of the Transformer model

3.4.2 LLMs

For comparison, we used models in which youko-8b was continually pre-trained on JParaCrawl v3.0. After the continual pre-training, we performed supervised fine-tuning in two ways: one using the WMT20 test set and other datasets, and the other using patent claims. These models served as baselines in our experiments.

3.4.3 Prompt

When using the prompt format for the inference described in Section 3.2 for the SFT training data, numbers that did not exist in the source sentences appeared at the beginning of the output sentences. Specific examples of this phenomenon are provided in Appendix B. While the exact cause of this issue remains unclear, this phenomenon occurs in Japanese-to-English translations regardless of the data used for the continual pre-training or SFT. Thus, it is hypothesized that this behavior may be attributable to the Japanese continual pre-training process of the youko-8b model. To determine if it is possible to suppress the occurrence of such extraneous numbers in the output, we conducted additional inference experiments by modifying the prompts to the format shown below.

> これを日本語から英語に翻訳してください。ただし文頭に関係のない数字を出さないようにしてください。:
>
> 日本語: `{Japanese_text}`
>
> 英語:

The English translation of the above prompt is:

> "Translate this from Japanese to English. However, do not start the sentence with an irrelevant number."

3.5 Investigation of Required Data Volume for Continual Pre-training

In this study, approximately 61 million sentence pairs of patent data were used for continual pre-training. To investigate how much data is necessary for effective continual pre-training, we saved checkpoints every 0.1 epoch (i.e., every 6.1M sentence pairs) during the training process. SFT was then applied to each of these intermediate checkpoints, and the translation performance was compared. For reference, the translation accuracy of the model where SFT was applied to youko-8b

without any continual pre-training is denoted as the result at "0 sentence pairs".

In addition to the original time-ordered data, we also experimented with two alternative data orderings: reversed chronological order and random order. The same procedure was applied to these variations to examine how the order of training data affects the effectiveness of continual pre-training.

3.6 Evaluation Metrics

For evaluation metrics, we employed BLEU (Papineni et al., 2002) and COMET (Rei et al., 2022). The BLEU scores were calculated using sacreBLEU (Post, 2018), whereas the COMET scores were obtained with the wmt22-comet-da model. Additionally, we analyzed win/lose cases by comparing the baseline translation results and the translation results of the LLM with the highest system-level scores, evaluating them at the sentence level for both BLEU and COMET.

4 Evaluation Results

4.1 Results of Continual Pre-training and SFT

Table 5 shows the translation accuracy achieved through the continual pre-training and SFT. The results indicate that models pre-trained with patent data significantly improved the BLEU scores compared with those pre-trained with JParaCrawl. Specifically, the BLEU score for the model pre-trained with patent data and fine-tuned with patent claims using full fine-tuning reached 50.7, compared with 43.5 for the model pre-trained with JParaCrawl. Similarly, LoRA fine-tuning achieved a BLEU score of 51.3 with patent data, significantly outperforming the 43.8 obtained with JParaCrawl. These results demonstrate that the continual pre-training on patent data effectively enables the model to acquire domain-specific knowledge.

Performing SFT with patent claims resulted in statistically significant improvements ($p < 0.05$) in the BLEU scores over the baseline model, achieving a BLEU score of 50.2. Among the SFT methods, LoRA achieved the highest BLEU score of 51.3, whereas full fine-tuning achieved 50.7. Although LoRA demonstrated superior BLEU scores, the COMET scores favored full fine-tuning, with values of 80.79 for LoRA and 81.25 for full fine-tuning.

When the inference prompt was improved, as described in Section 3.4.3, both the BLEU and COMET scores increased across all SFT methods. After prompt improvements, the BLEU score for LoRA increased to 52.3, and that of full fine-tuning improved to 52.0. Similarly, the COMET scores increased to 82.52 for LoRA and 82.55 for full fine-tuning. The analysis of the outputs revealed that the improved prompt successfully eliminated extraneous numbers at the beginning of sentences, which contributed positively to the translation quality. Examples of outputs before and after prompt modification are provided in Appendix B.

As an additional experiment, we randomly selected 100 test samples and translated them using GPT-4o to compare its performance with the proposed method. The GPT-4o translation was conducted under two conditions: (1) **Zero-shot Translation**, where the model was prompted to generate translations without any additional context, and (2) **Three-shot Translation**, where three example translations were randomly selected from the SFT training data and provided as in-context examples for few-shot translation. This comparison was conducted to provide a reference point for the translation accuracy of commercially available LLMs. Given the results of WMT23, where GPT-4 demonstrated superior translation performance compared to existing models, we aimed to assess how well GPT-4o performs specifically on patent claims. Additionally, we investigated the extent to which its performance improves with a 3-shot prompt and how our proposed approach compares to it. The translation results were evaluated using BLEU and COMET scores and compared against both the baseline and the model that achieved the highest translation accuracy in Table 5, which is referred to as the *Proposed Method* and shown in Table 6. As a result, in terms of BLEU, even with three-shot translation, GPT-4o exhibited a statistically significant drop in scores compared to both the baseline and the proposed method. However, in terms of COMET, no such trend was observed, and the difference was not statistically significant.

4.2 Required Data Volume for Continual Pre-training

4.2.1 Quantitative Evaluation

The BLEU and COMET scores for each data ordering (time-ordered, reversed-order, and random) are compared in Figures 2 and 3, respec-

Training Method	BLEU	COMET
baseline model	50.2	81.92
Continual Pre-training + SFT (Method)		
JParaCrawl + WMT (Full)	38.0	81.42
JParaCrawl + WMT (LoRA)	34.2	80.70
JParaCrawl + patent claims (full)	43.5	81.36
JParaCrawl + patent claims (LoRA)	43.8	81.37
patent + patent claims (full)	50.7∗	81.25
patent + patent claims (LoRA)	51.3∗	80.79
patent + patent claims (full) + prompt improvement	52.0∗	**82.55**∗
patent + patent claims (LoRA) + prompt improvement	**52.3**∗	82.52∗

Table 5: BLEU and COMET scores for each training method. ∗ indicates a significant difference from the baseline ($p < 0.05$).

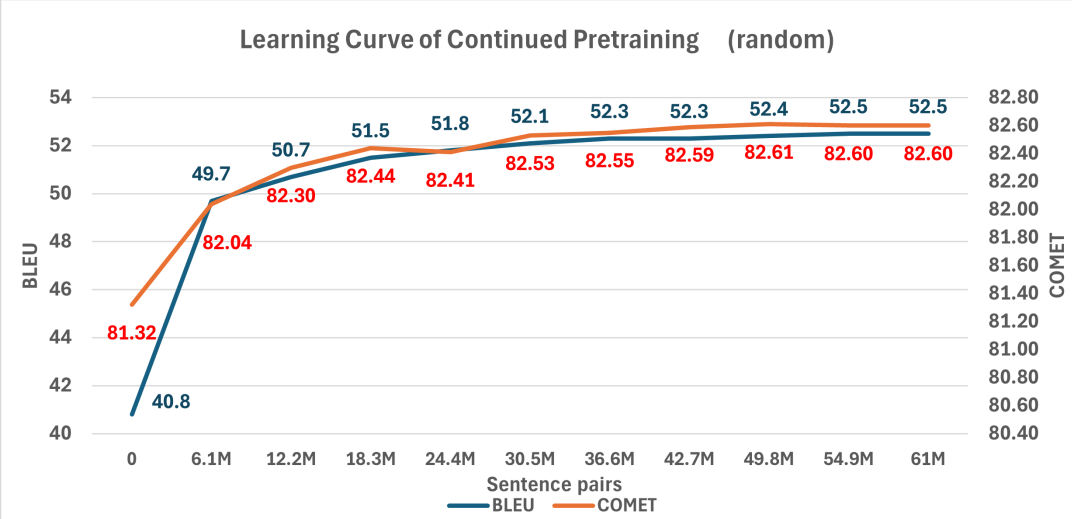


Figure 1: Learning Curve of Continual Pre-training (random)

Models	BLEU	COMET
baseline (Transformer enc-dec)	54.5∗	0.8296
proposed method	59.3∗	0.8345
GPT-4o (zero-shot)	44.8	0.8282
GPT-4o (three-shot)	48.2	0.8324

Table 6: Comparison of translation by GPT-4o. ∗ indicates a significant difference from the GPT-4o (Three-shot) ($p < 0.05$).

tively. Based on these two figures, the randomly ordered data yielded the best overall translation performance. Figure 1 shows the translation evaluation results when supervised fine-tuning (SFT) was conducted at every 6.1M sentence pairs using the randomly ordered data. For reference, the full results and specific values for the time-ordered and reversed-order settings are provided in Appendix C.

At "0 sentence pairs", i.e., where the base model (youko-8b) was directly fine-tuned with patent claim data, the BLEU score was 40.8. However, by 6.1M sentence pairs, the BLEU score had increased to 49.7, demonstrating that even a small amount of data significantly improved translation accuracy through the continual pre-training.

BLEU and COMET scores showed a substantial increase up to 30.5M sentence pairs, achieving approximately 90% of the total performance gain observed. Beyond this point, BLEU and COMET scores continued to rise, albeit more gradually.

4.2.2 Qualitative Evaluation

As a qualitative evaluation, we compared the translation results of the models subjected to SFT at various stages: before the continual pre-training, and at 24.4M sentence pairs, 42.7M sentence pairs, and 61M sentence pairs of continual pre-training. Specific examples are presented in Table 8. These examples demonstrate significant improvements in translation quality after the continual pre-training compared with that be-

COMET difference	#cases	by Baseline	by LLM
0.1 to 0.2, LLM win	613	0.5728	0.9756
0.1 to 0.2, LLM lose	355	0.9802	0.7785
0.2 or higher, LLM win	203	0.3853	0.9864
0.2 or higher, LLM lose	380	0.7320	0.2618

Table 7: Median Sentence Length Ratios classified by COMET Score Differences and Win/Lose Cases

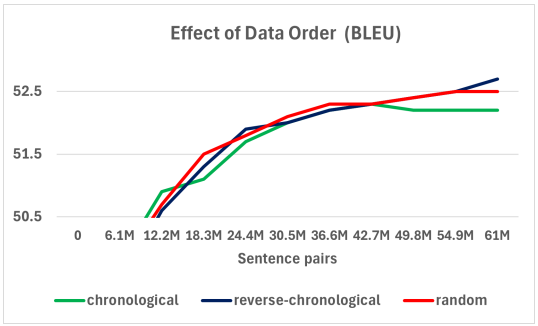


Figure 2: Effect of Data Order (BLEU)

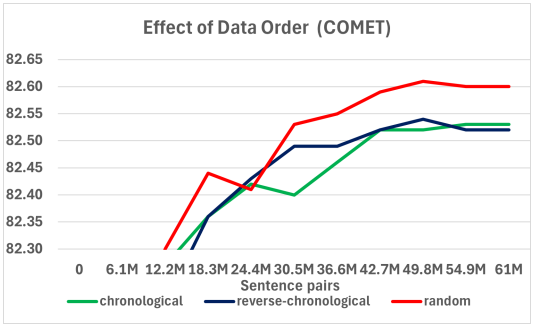


Figure 3: Effect of Data Order (COMET)

fore. This results indicates that the model acquired knowledge related to patents and parallel translations through the continual pre-training. Although the differences between the 24.4M sentence pairs and the completion of continual pre-training appeared minor in this example, variations in expression were observed, and a statistically significant improvement in the sentence BLEU scores was confirmed.

4.3 Analysis

To analyze the tendencies, sentence-level COMET and BLEU scores were calculated, and win/lose sentence sets were divided accordingly. To simplify the analysis and avoid difficulties caused by minor differences, examples with COMET differences between 0.1 and 0.2 (win/lose) and those with differences of 0.2 or higher (win/lose) were extracted, with 50 examples selected for each category. The total number of cases in which these differences occurred, along with the length ratio of the reference sentences to the translation results (baseline and LLM) for the selected 50 examples, is presented in Table 7. The length ratio between the reference and the translated sentences was calculated as a measure because many lose cases showed omissions in the translations, as observed in the selected examples.

The analysis confirmed that, as observed during the initial evaluation, translation outputs in the lose cases generally exhibited a lower length ratio than the reference sentences. This finding indicates that omissions occurred more frequently in the lose cases.

Additionally, manual inspection was conducted to provide a more detailed analysis of the specific errors in translations generated by the baseline model and the LLM.

Based on the manual inspection, among the cases where the LLM outperformed the baseline, 32 out of 50 examples exhibited omissions in the baseline translation, 5 examples showed repetition, and 13 examples contained both omissions and repetitions. Conversely, in cases where the LLM underperformed, 38 out of 50 examples exhibited omissions, 7 examples exhibited both hallucinations and omissions, and 5 examples exhibited repetition.

For example, in one case where the LLM outperformed the baseline, the source sentence described a semi-aromatic polyamide resin including multiple chemical conditions and formula-based constraints. The baseline translation retained only the formulas, such as “10 eq/t AEG+CEG 140 eq/t,” while omitting the entire description of the resin structure. In contrast, the LLM output correctly preserved the chemical structure, including “a structural unit obtained from hexamethylenediamine and terephthalic acid,” and maintained the constraints, indicating a more faithful translation.

In another representative case, the baseline output included severe repetition of the phrase “cantilever shaped” over 60 times, resulting in a clearly failed translation. The LLM translation avoided this repetition entirely, outputting a coherent description such as “with at least one cantilevered

source sentence

各生物学的成分がヌクレオチド配列または微生物株のうちの少なくとも 1 つである、請求項 1 から 11 のいずれか一項に記載のシステム。遺伝子改変を組み込んだ少なくとも 1 つの目的の産物の遺伝子製造システムにおける産生を制御するためにビルドグラフデータ構造を生成するためのコンピュータ実装方法であって、 生物学的ワークフローの記述であって、生物学的成分の表現を含む記述にアクセスすることと、 前記ワークフロー記述に少なくとも一部は基づいて、ビルドグラフデータ構造をアセンブルすることと を含み、 前記ビルドグラフデータ構造内で、各生物学的成分が、複数のレベルのうちのあるレベルにあるノードによって表され、 前記複数のレベルのうちの所与のレベルにおける 1 つまたは複数のソースノード、および前記所与のレベルの子レベルにおける宛先ノードが、前記子レベルに対応する、1 つまたは複数の反応グループのうちのある反応グループを構成し、 各反応グループが、前記子レベルにおける前記反応グループの前記宛先ノードによって表される生物学的成分を産生するために、前記所与のレベルにおける前記 1 つまたは複数のソースノードによってそれ自体が表される 1 つまたは複数の生物学的成分間の反応を表し、 前記子レベルにおける 1 つまたは複数の宛先ノードが、前記所与のレベルの孫レベルにおける 1 つまたは複数の反応グループのうちのある反応グループ内の 1 つまたは複数のソースノードとして作用し、 前記複数のレベルのうちの最終レベルにおける少なくとも 1 つの宛先ノードが、異なるレベルにおける生物学的成分間の反応によって引き起こされた遺伝子改変を組み込んだ、前記少なくとも 1 つの目的の産物を表し、 前記ビルドグラフデータ構造を処理することにより、前記少なくとも 1 つの目的の産物が産生される、方法。前記少なくとも 1 つの目的の産物がヌクレオチド配列または微生物株を含む、請求項 13 に記載の方法。

reference translation

access a description of a biological workflow, wherein the description includes representations of biological components; and assemble a build graph data structure based at least in part upon the workflow description, wherein, in the build graph data structure, each biological component is represented by a node that resides at a level of a plurality of levels, wherein one or more source nodes, at a given level of the plurality of levels, and a destination node, at a child level of the given level, constitute a reaction group of one or more reaction groups corresponding to the child level, wherein each reaction group represents a reaction between one or more biological components that are themselves represented by the one or more source nodes at the given level, to produce a biological component represented by the destination node of the reaction group at the child level, wherein one or more destination nodes at the child level act as one or more source nodes in a reaction group of one or more reaction groups at a grandchild level of the given level, and wherein at least one destination node at a final level of the plurality of levels represents the at least one product of interest, which incorporates genetic modifications caused by reactions among biological components at different levels, and wherein processing the build graph data structure results in production of the at least one product of interest.

before continual pre-training (BLEU=0.001, COMET=50.56)

The method of claim 13, wherein the at least one product of interest comprises a nucleotide sequence or a microbial strain.

after continual pre-training with 24.4M sentence pairs (BLEU=68.04, COMET=86.63)

A computer-implemented method for generating a build graph data structure to control production in a genetic manufacturing system of at least one product of interest incorporating a genetic modification, the method comprising: accessing a description of a biological workflow, the description comprising a representation of biological components; and assembling a build graph data structure based at least in part on the workflow description, wherein within the build graph data structure, each biological component is represented by a node at a level of a plurality of levels, and wherein one or more source nodes at a given level of the plurality of levels and a destination node at a child level of the given level constitute a reaction group of one or more reaction groups corresponding to the child level, and wherein each reaction group represents a reaction between one or more biological components represented by the one or more source nodes at the given level to produce a biological component represented by the destination node of the reaction group at the child level, and wherein the one or more destination nodes at the child level act as one or more source nodes in a reaction group at one or more reaction groups at a grandchild level of the given level, and wherein at least one destination node at a final level of the plurality of levels represents the at least one product of interest incorporating a genetic modification caused by a reaction between biological components at different levels, and wherein the at least one product of interest is produced by processing the build graph data structure.

after continual pre-training (with 42.7M sentence pairs) (BLEU=68.94, COMET=86.77)

A computer-implemented method for generating a build graph data structure to control production in a genetic manufacturing system of at least one product of interest incorporating a genetic modification, the method comprising: accessing a description of a biological workflow, the description comprising a representation of biological components; and assembling a build graph data structure based at least in part on the workflow description, wherein within the build graph data structure, each biological component is represented by a node at a level of a plurality of levels, and wherein one or more source nodes at a given level of the plurality of levels and a destination node at a child level of the given level constitute a reaction group of one or more reaction groups corresponding to the child level, and wherein each reaction group represents a reaction between one or more biological components represented by the one or more source nodes at the given level to produce a biological component represented by the destination node of the reaction group at the child level, and wherein the one or more destination nodes at the child level act as one or more source nodes in a reaction group of one or more reaction groups at a grandchild level of the given level, and wherein at least one destination node at a final level of the plurality of levels represents the at least one product of interest incorporating a genetic modification caused by a reaction between biological components at different levels, and wherein the at least one product of interest is produced by processing the build graph data structure.

after completing continual pre-training (with 61M sentence pairs) (BLEU=70.75, COMET=86.8)

A computer-implemented method for generating a build graph data structure to control production of at least one product of interest in a genetic manufacturing system incorporating genetic modifications, the method comprising: accessing a description of a biological workflow, the description comprising a representation of biological components; and assembling a build graph data structure based at least in part on the workflow description, wherein within the build graph data structure, each biological component is represented by a node at a level of a plurality of levels, and wherein one or more source nodes at a given level of the plurality of levels and a destination node at a child level of the given level constitute a reaction group of one or more reaction groups corresponding to the child level, and wherein each reaction group represents a reaction between one or more biological components represented by the one or more source nodes at the given level to produce a biological component represented by the destination node of the reaction group at the child level, and wherein the one or more destination nodes at the child level act as one or more source nodes in a reaction group of one or more reaction groups at a grandchild level of the given level, and wherein at least one destination node at a final level of the plurality of levels represents the at least one product of interest incorporating genetic modifications caused by reactions between biological components at different levels, and wherein the at least one product of interest is produced by processing the build graph data structure.

Table 8: Example (1): Improvements through Continual Pre-training

beam.” This suggests that the LLM reduced unnecessary repetition, contributing to the improved translation scores.

However, when the LLM underperformed, different issues arose. In one example, the source sentence defined a chemical compound using a formula (I) and a detailed list of structural groups such as “X is C1–C6 alkyl...”, “R1 is a halo...”, and “Ar is an aryl or heteroaryl group.” While the baseline output covered all these elements almost verbatim, the LLM output stopped at “A compound of formula (I)...,” omitting all detailed structural components that followed.

A more extreme example of repetition was observed in a case involving a list of agents used to induce a stress response. The original sentence listed items from a) to y), including phrases like “interferon gamma,” “poly(IC),” and “monophosphoryl lipid A.” The baseline correctly stopped at item p) or so. In contrast, the LLM continued well beyond the source list, generating items labeled “z), aa), bb), ... lll),” all filled with repeated phrases like “lipooligosaccharide isolated from gram positive bacteria.” This artificial extension of the list demonstrates a severe repetition pattern unique to LLMs.

Although specific examples are not cited in detail here, it was also observed that in some LLM outputs, lists of detailed items were occasionally collapsed into a single concept. For instance, when the source sentence enumerated specific cancer types, the LLM sometimes generalized this into “cancer” rather than preserving individual names. This abstraction behavior, while possibly acceptable in some domains, represents a unique challenge in the accurate translation of patent claims that demand precision.

For the full outputs corresponding to the examples above, please refer to Appendix A.

5 Conclusion

This study investigated the effectiveness of LLMs for patent claim translation through the application of continual pre-training and SFT with domain-specific parallel data. The results demonstrated that LLMs, fine-tuned with patent-specific datasets, outperformed traditional Transformer encoder-decoder based models in terms of BLEU and COMET scores, thereby highlighting their superior ability to handle the intricate sentence structures and technical terminology characteristic of patent documents. A notable improvement was observed in the reduction of common translation issues such as omissions and repetitions, highlighting the capacity of LLMs to better retain and reproduce the detailed content of the source text. Furthermore, The experimental findings underscore the critical role of prompt design in enhancing translation performance, as improved prompts led to more accurate results. The study further showed the impact of data volume on the continual pre-training, indicating that substantial enhancements in translation performance can be achieved with relatively moderate data sizes. These findings provide a strong foundation for the potential of LLMs as a viable tool for high-quality patent translation tasks, contributing to advancements in the field of specialized machine translation.

References

F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang. 2022. Language-agnostic BERT sentence embedding. In *Proc. 60th ACL*, pages 878–891.

M. Fuji, A. Fujita, M. Utiyama, E. Sumita, and Y. Matsumoto. 2015. Patent claim translation based on sublanguage-specific sentence structure. In *Proc. Machine Translation Summit XV: Papers*, pages 1–16.

A. Fujii, M. Utiyama, M. Yamamoto, T. Utsuro, T. Ehara, H. Echizen-ya, and S. Shimohata. 2008. Overview of the Patent Translation Task at the NTCIR-7 Workshop. In *Proc. NTCIR*, pages 389–400.

I. Goto, Ka-Po Chow, B. Lu, E. Sumita, and Benjamin K Tsou. 2013. Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. In *Proc. NTCIR*, pages 260–286.

J. Guo, H. Yang, Z. Li, D Wei, H. Shang, and X. Chen. 2024. A novel paradigm boosting translation capabilities of large language models. In *Proc. NAACL 2024*, pages 639–649.

E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proc. 10th ICLR*, pages 1–13.

J. van der Linde G. Ramírez-Sánchez K. Heafield, E. Farrow and D. Wiggins. 2022. The EuroPat Corpus: A parallel corpus of european patent data. In *Proc. 13th LREC*, pages 732–740.

T. Kocmi et al. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proc. 7th WMT*, pages 1–45.

T. Kocmi et al. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *In Proc. of 8th WMT*, pages 1–42.

T. Kocmi et al. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proc. 9th WMT*, pages 1–46.

M. Kondo, T. Utsuro, and M. Nagata. 2024. Enhancing translation accuracy of large language models through continual pre-training on parallel data. In *Proc. 21th IWSLT*, pages 203–220.

T. Kudo and J. Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proc. EMNLP*, pages 66–71.

A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, pages 1–8.

Y. Moslem, R. Haque, J.D. Kelleher, and A. Way. 2023. Adaptive machine translation with large language models. In *Proc. 24th EAMT*, pages 227–237. European Association for Machine Translation.

M. Nagata, M. Morishita, K. Chousa, and N. Yasuda. 2024. JaParaPat: A large-scale Japanese-English parallel patent application corpus. In *Proc. 15th LREC*, pages 9452–9462.

T. Nakazawa, C. Ding, H. Mino, I. Goto, G. Neubig, and S. Kurohashi. 2016. Overview of the 3rd Workshop on Asian Translation. In *Proc. the 3rd Workshop on Asian Translation (WAT2016)*, pages 1–46.

M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proc. NAACL*, pages 48–53.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proc. 40th ACL*, pages 311–318.

M. Post. 2018. A call for clarity in reporting BLEU scores. In *Proc. 3rd WMT*, pages 186–191.

J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He. 2023. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proc. 29th ACM SIGOPS*, pages 3505–3506.

R. Rei, J. C. de Souza, D. Alves, C. Zerva, A. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proc. 7th WMT*, pages 578–585.

NLLB Team et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv*, 2207.04672:1–192.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Proc. 31st NIPS*, pages 1–11.

H. Xu, K. Young Jin, S. Amr, and A. Hany Hassan. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *Proc. 12th ICLR*, pages 1–21.

X. Zhang, N. Rajabi, K. Duh, and P. Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proc. 8th WMT*, pages 468–481.

J. Zheng, H. Hong, F. Liu, X. Wang, J. Su, Y. Liang, and S. Wu. 2024. Fine-tuning Large Language Models for Domain-specific Machine Translation.

A Example for Analysis

This appendix provides detailed examples of translation outputs referenced in the section 4.3.

example 2-1 (case of omission)

source sentence

ヘキサメチレンジアミンとテレフタル酸から得られる構成単位、及び 11 −アミノウンデカン酸又はウンデカンラクタムから得られる構成単位を含有し、相対粘度 (RV) が式 (1) の範囲であり、アミノ基末端濃度 (AEG) 、 カルボキシ基末端濃度 (CEG) 及びモノカルボン酸でアミノ基末端を封鎖した末端濃度 (EC) の関係が式 (2) 及び (3) を満たす半芳香族ポリアミド樹脂。1.95 ≦ RV ≦ 3.50 ・ ・ (1) 10eq/t ≦ AEG+CEG ≦ 140eq/t ・ ・ (2) (AEG+CEG) / (AEG+CEG+EC) ≦ 0.50 ・ ・ (3)

reference translation

wherein the resin contains a constituent unit obtained from hexamethylenediamine and terephthalic acid and a constituent unit obtained from 11-aminoundecanoic acid or undecane lactam, wherein a relative viscosity (RV) of the semi-aromatic polyamide resin satisfies the following formula (I): 1.95RV3.50, and wherein a relationship among a concentration of terminal amino groups (AEG), a concentration of terminal carboxyl groups (CEG) and a concentration of terminal amino groups blocked by a monocarboxylic acid (EC) satisfies the following formula (II): 10 eq/tAEG+CEG140 eq/t, and the following formula (III): (AEG+CEG)/(AEG+CEG+EC)0.50.

baseline translation (BLEU=1.8, COMET=40.49)

10 eq/tAEG+CEG140 eq/t (2)(AEG+CEG)/(AEG+CEG+EC)0.50 (3)

LLM translation (BLEU=26.6, COMET=73.74)

A semi-aromatic polyamide resin comprising a structural unit obtained from hexamethylenediamine and terephthalic acid and a structural unit obtained from 11-aminoundecanoic acid or undecane lactam, wherein the semi-aromatic polyamide resin has a relative viscosity (RV) in a range of formula (1), and a relationship between an amino group terminal concentration (AEG), a carboxy group terminal concentration (CEG), and a terminal concentration (EC) obtained by blocking an amino group terminal with a monocarboxylic acid satisfies formulas (2) and (3):

example 2-2 (case of repetition)

source sentence

前記流路遮断バルブは、 内部に前記閉鎖部材を収容し、カンチレバー形状からなる少なくとも 1 つの片持ちばりを備え、円筒形に形成されて、前記連通流路の流入口に挿入されるように設置されるボディー; 及び、 一側は前記片持ちばりから突出形成される係止部により支持され、他側は前記閉鎖部材と接触するように設置されるリング部材; を含み、 前記リング部材は設定された温度以上になると、前記閉鎖部材が中心部を通過するように変形されて、前記閉鎖部材を前記連通流路の内部に向けて移動させる、 ことを特徴とする、請求項 17 に記載のバルブアセンブリ。

reference translation

The valve assembly of claim 17, wherein the flow path blocking valve includes:a body for containing the blocking member therein, and providing at least one cantilever portion formed in a cantilever shape, the body formed in a cylindrical shape and disposed to be inserted into the inlet of the communication flow path; anda ring member having one side supported by a locking portion protruding from the cantilever portion and the other side disposed to contact the blocking member, wherein the ring member moves the blocking member towards the inside of the communication flow path by deforming the blocking member to pass through a central part of the ring member when the internal temperature exceeds the preset temperature.

baseline translation (BLEU=7.7, COMET=34.42)

The valve assembly of claim 17, wherein the flow path shut-off valve comprises:a body configured to accommodate the closing member therein, having at least one cantilever shaped

LLM translation (BLEU=40.5, COMET=82.57)

The valve assembly according to claim 17, wherein the flow path blocking valve comprises:a body, which is formed in a cylindrical shape and is provided with at least one cantilevered beam, which is formed in a cantilever shape and accommodates the closing member therein, and is inserted into the inlet of the communication flow path; anda ring member, one side of which is supported by a locking portion protruding from the cantilevered beam, and the other side of which is in contact with the closing member, wherein the ring member is deformed to allow the closing member to pass through the center portion thereof when the temperature is equal to or higher than a set temperature, so that the closing member is moved toward the inside of the communication flow path.

Table 9: Example (2): LLM Outperforming Baseline

Table 9 presents representative examples in which the LLM-generated translations significantly outperformed those of the baseline model. These include instances where the baseline output exhibits severe omissions (e.g., outputting only numerical expressions and omitting structural descriptions) or extreme repetition (e.g., repeating the phrase “cantilever shaped” more than 60 times). In contrast, the LLM translations tend to preserve semantic structure more faithfully and exhibit fewer surface-level artifacts. BLEU and COMET scores are provided for each translation to support the qualitative analysis.

example 3-1 (case of omission)

source sentence

式 (I) : (式中、X は、C 1 〜C 6 アルキル又は C 2 〜C 6 アルケニルであり; Y は、 であり; W は、O 又は S であり; R 2 は、H、アルキル又はアルケニルであり; Z は、ヘテロシクロアルキル又はヘテロアリール基であり、ここで、ヘテロシクロアルキル又はヘテロアリール基は任意選択的に置換されており; R 1 は、ハロ、シクロアルキル、ヘテロシクロアルキル、アリール又はヘテロアリール基であり、ここで、シクロアルキル、ヘテロシクロアルキル、アリール又はヘテロアリール基は任意選択的に置換されており; Ar は、アリール又はヘテロアリール基である) の化合物又はその薬学的に許容可能な塩若しくはプロドラッグ。

reference translation

A compound of formula (I):or a pharmaceutically acceptable salt or prodrug thereof, wherein:X is C1-C6 alkyl or C2-C6 alkenyl;Y isW is O or S;R2 is H, alkyl or alkenyl;Z is heterocycloalkyl or heteroaryl group, which heterocycloalkyl or heteroaryl group is optionally substituted;R1 is a halo, cycloalkyl, heterocycloalkyl, aryl or heteroaryl group, which cycloalkyl, heterocycloalkyl, aryl or heteroaryl group is optionally substituted;Ar is an aryl or heteroaryl group.

baseline translation (BLEU=74.97, COMET=87.9)

A compound of Formula (I):or a pharmaceutically acceptable salt or prodrug thereof, whereinX is C1-C6 alkyl or C2-C6 alkenyl;Y isW is O or S;R2 is H, alkyl, or alkenyl;Z is a heterocycloalkyl or heteroaryl group, wherein the heterocycloalkyl or heteroaryl group is optionally substituted;R1 is a halo, cycloalkyl, heterocycloalkyl, aryl, or heteroaryl group, wherein the cycloalkyl, heterocycloalkyl, aryl, or heteroaryl group is optionally substituted; andAr is an aryl or heteroaryl group.

LLM translation (BLEU=1.64, COMET=42.52)

A compound of formula (I):or a pharmaceutically acceptable salt or prodrug thereof, wherein:

example 3-2 (case of repetition)

source sentence

ストレス応答を誘導することができる前記作用物質が、a) インターフェロンガンマ;b)IVIG;c) 単球条件培地;d) 好中球細胞外トラップに曝露された末梢血単核球からの上清;e) 単球との共培養物;f)IVIG で前処置された単球との共培養物;g)T 細胞との共培養物;h)T 細胞刺激に曝露された T 細胞との共培養物;i)NK 細胞との共培養物;j) グラム陽性菌から単離されたペプチドグリカン;k) グラム陽性菌から単離されたリポテイコ酸;l) グラム陽性菌から単離されたリポタンパク質;m) マイコバクテリアから単離されたリポアラビノマンナン、n) 酵母細胞ウェルから単離されたザイモサン;o) ポリアデニル酸－ポリウリジル酸;p) ポリ (IC) ; q) リポ多糖;r) モノホスホリルリピド A;s) フラジェリン;t) ガーディキモド;u) イミキモド;v)R848;w)CpG モチーフを含むオリゴヌクレオシド;x)23S リボソーム RNA; 及び y) それらの組み合わせからなる群より選択される、請求項 131 又は 136 に記載の方法。

reference translation

The method of claim 131 or 136, wherein said agent capable of inducing a stress response is selected from the group consisting of a) interferon gamma; b) IVIG; c) monocyte conditioned media; d) supernatant from neutrophil extracellular trap exposed peripheral blood mononuclear cells; e) co-culture with monocytes; f) co-culture with monocytes that have been pretreated with IVIG; g) co-culture with T cells; h) co-culture with T cells that have been exposed to a T cell stimulus; i) co-culture with NK cells; j) peptidoglycan isolated from gram positive bacteria; k) lipoteichoic acid isolated from gram positive bacteria; l) lipoprotein isolated from gram positive bacteria; m) lipoarabinomannan isolated from mycobacteria, n) zymosan isolated from yeast cell well; o) Polyadenylic-polyuridylic acid; p) poly (IC); q) lipopolysaccharide; r) monophosphoryl lipid A; s) flagellin; t) Gardiquimod; u) Imiquimod; v) R848; w) oligonucleosides containing CpG motifs; x) 23S ribosomal RNA; and y) a combination thereof.

baseline translation (BLEU=55.27, COMET=84.2)

The method of claim 131 or 136, wherein the agent capable of inducing a stress response is selected from the group consisting of: a) interferon gamma; b) IVIG; c) monocyte conditioned medium; d) supernatant from peripheral blood mononuclear cells exposed to a neutrophil extracellular trap; e) co-culture with monocytes; f) co-culture with monocytes pretreated with IVIG; g) co-culture with T cells; h) co-culture with T cells exposed to a T cell stimulus; i) co-culture with NK cells; j) peptidoglycan isolated from Gram-positive bacteria; k) lipoteichoic acid isolated from Gram-positive bacteria; l) lipoprotein isolated from Gram-positive bacteria; m) lipoarabinomannan isolated from mycobacteria, n) zymosan isolated from yeast cell wells; o) polyadenylate-polyuridylate; p) poly(IC);

LLM translation (BLEU=9.37, COMET=6.72)

The method of claim 131 or 136, wherein the agent capable of inducing a stress response is selected from the group consisting of: a) interferon gamma; b) IVIG; c) monocyte conditioned media; d) supernatant from peripheral blood mononuclear cells exposed to neutrophil extracellular trap; e) co-culture with monocytes; f) co-culture with monocytes pre-treated with IVIG; g) co-culture with T cells; h) co-culture with T cells exposed to T cell stimulation; i) co-culture with NK cells; j) peptidoglycan isolated from gram positive bacteria; k) lipoteichoic acid isolated from gram positive bacteria; l) lipoprotein isolated from gram positive bacteria; m) lipooligosaccharide isolated from gram positive bacteria; n) lipooligosaccharide isolated from gram positive bacteria; o) lipooligosaccharide isolated from gram positive bacteria; p) lipooligosaccharide isolated from gram positive bacteria; q) lipooligosaccharide isolated from gram positive bacteria; r) lipooligosaccharide isolated from gram positive bacteria; s) lipooligosaccharide isolated from gram positive bacteria; t) lipooligosaccharide isolated from gram positive bacteria; u) lipooligosaccharide isolated from gram positive bacteria; v) lipooligosaccharide isolated from gram positive bacteria; w) lipooligosaccharide isolated from gram positive bacteria; x) lipooligosaccharide isolated from gram positive bacteria; y) lipooligosaccharide isolated from gram positive bacteria; z) lipooligosaccharide isolated from gram positive bacteria; aa) lipooligosaccharide isolated from gram positive bacteria; · · · iii) lipooligosaccharide isolated from gram positive bacteria; jjj) lipooligosaccharide isolated from gram positive bacteria; kkk) lipooligosaccharide isolated from gram positive bacteria; lll) lipooligosacchar

Table 10: Example (3): LLM Underperforming Baseline

Table 10 contains examples where the LLM underperformed compared to the baseline. These include cases of significant omissions (e.g., truncating a list of chemical constituents) and repetition (e.g., extending a list of items beyond what was present in the source text by repeatedly generating the same phrase, such as “ z), aa), bb), ... lll) ”). These examples illustrate types of degradation unique to LLM outputs, particularly in structured or enumerative patent language. As in the previous table, evaluation scores are provided alongside each translation.

B Prompt

Table 11 shows the output examples from the prompt described in Section 3.2 as well as the improved prompt described in Section 3.4.3. The numbers that appear tend to correspond to the subsequent number following those present in the source text.

source sentence
細胞の単位用量が、規定の数の CD8 + /CCR7 + 細胞、CD4 + /CCR7 + 細胞、CD8 + /CD27 + 細胞、CD4 + /CD27 + 細胞、CD8 + /CCR7 + /CD27 + 細胞、および/または CD4 + /CCR7 + /CD27 + 細胞を含む、請求項 113 記載の方法。
reference translation
The method of claim 113, wherein the unit dose of cells comprises a defined number of CD8+/CCR7+ cells, CD4+/CCR7+ cells, CD8+/CD27+ cells, CD4+/CD27+ cells, CD8+/CCR7+/CD27+ cells and/or CD4+/CCR7+/CD27+ cells.
translation by LLM
114. The method of claim 113, wherein the unit dose of cells comprises a defined number of CD8+/CCR7+ cells, CD4+/CCR7+ cells, CD8+/CD27+ cells, CD4+/CD27+ cells, CD8+/CCR7+/CD27+ cells, and/or CD4+/CCR7+/CD27+ cells.
translation by LLM (with improved prompt)
The method of claim 113, wherein the unit dose of cells comprises a defined number of CD8+/CCR7+ cells, CD4+/CCR7+ cells, CD8+/CD27+ cells, CD4+/CD27+ cells, CD8+/CCR7+/CD27+ cells, and/or CD4+/CCR7+/CD27+ cells.

Table 11: Translations Generated by the Prompt in Section 3.2

C Translation Evaluation Results by Data Order (Time-Ordered, Reversed-Order)

This appendix presents the detailed translation evaluation results for the different data orderings—time-ordered and reversed-order—used during the continual pre-training. For each of these settings, both the BLEU and COMET scores are provided. The figure includes a comparison of these scores, highlighting the differences in translation performance across the various data arrangements.

The results for the time-ordered and reversed-order configurations are shown in Figures 4 and 5.

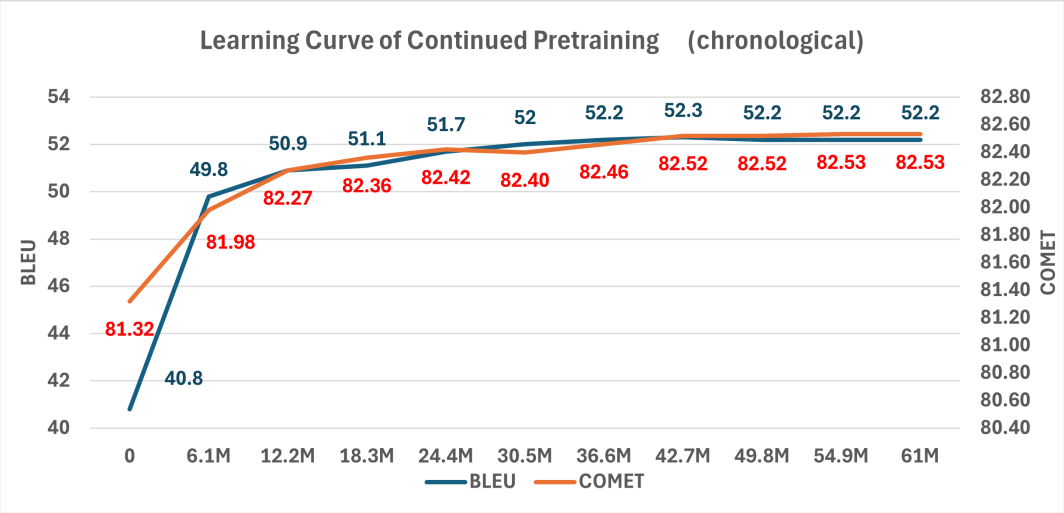


Figure 4: Learning Curve of Continual Pre-training

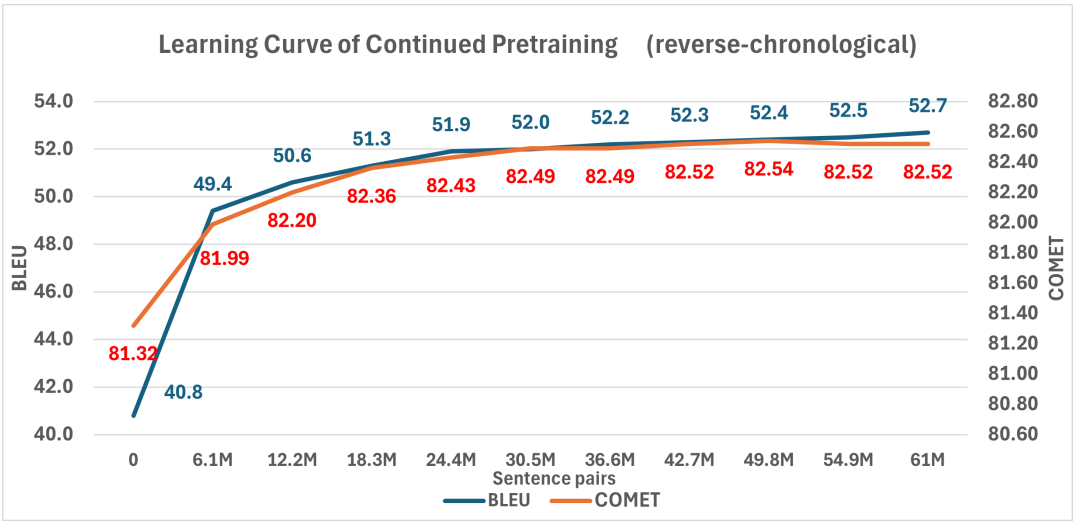


Figure 5: Learning Curve of Continual Pre-training (reverse)

D Sustainability Statement

D.1 CO2 Emission Related to Experiments

Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.432 kgCO_2eq/kWh. A cumulative of 400 hours of computation was performed on hardware of type A100 SXM4 80 GB (TDP of 400W).

Total emissions are estimated to be 34.56 kgCO_2eq of which 0 percents were directly offset.

Estimations were conducted using the MachineLearning Impact calculator presented in Lacoste et al. (2019).

The Devil is in the Details: Assessing the Effects of Machine-Translation on LLM Performance in Domain-Specific Texts

Javier Osorio[1], Afraa Alshammari[2], Naif Alatrush[2], Dagmar Heintze[2], Amber Converse[1], Sultan Alsarra[3], Latifur Khan[2], Patrick T. Brandt[2], Vito D'Orazio[4]

[1]University of Arizona [2]University of Texas at Dallas
[3]King Saud University [4]West Virginia University

Correspondence: josorio1@arizona.edu

Abstract

Conflict scholars increasingly use computational tools to track violence and cooperation at a global scale. To study foreign locations, researchers often use machine translation (MT) tools, but rarely evaluate the quality of the MT output or its effects on Large Language Model (LLM) performance. Using a domain-specific multilingual parallel corpus, this study evaluates the quality of several MT tools for text in English, Arabic, and Spanish. Using ConfliBERT, a domain-specific LLM, the study evaluates the effect of MT texts on model performance and finds that MT texts tend to yield better results than native-speaker written texts. The MT quality assessment reveals considerable translationese effects in vocabulary reduction, loss of text specialization, and syntactical changes. Regression analysis at the sentence level reveals that such distortions, particularly reductions in general and domain vocabulary rarity, artificially boost LLM performance by simplifying the MT output. This finding cautions researchers about uncritically relying on MT without considering MT-induced data loss.

1 Introduction

Political scientists, like many other domain-specific users, often rely on computational tools to make sense of large volumes of data. In particular, conflict scholars increasingly use computational methods to analyze global dynamics of political conflict and cooperation in foreign locations. To do so, researchers frequently rely on machine translations (MT) to translate political text from different languages (Boschee et al., 2018; Halterman et al., 2023). Despite the growing research on MT quality (Liu and Zhu, 2023; Kahlon and Singh, 2023; Lee, 2023; Ahrenberg, 2017), social scientists seldom evaluate the quality of the MT output nor its consequences on model performance. Careful researchers may be concerned about MT quality due to data loss or incorrect translations, particularly for low-resource languages (De Vries et al., 2018; Licht et al., 2024; Bartaškevičius, 2024) or specialized domains requiring precise terminology (Cambedda et al., 2021). MT-induced changes to the source text, known as translationese effect (Gellerstam, 1986), may result in considerable alterations of the output text, making translationese especially crucial to investigate in domain-specific translations where seemingly minor distortions of the output text may lead to incorrect inference. Moreover, there is little work analyzing the impact of MT quality on Large Language Model (LLM) performance (Huang and Liu, 2024). Consequently, the quality of the MT text often gets overlooked, and its effects on LLM performance remain ignored. For researchers tracking conflict around the world, disregarding translationese or its effects on LLM performance may lead to missing important signals about security threats or cooperation.

By using a multilingual parallel corpus from the United Nations (Ziemski et al., 2016), this study analyzes the quality of various MT tools for English, Arabic, and Spanish and evaluates the effects of MT distortions on LLM performance on tasks related to political conflict and cooperation. In particular, the study evaluates four MT tools, Google Translate (GT) (Google Cloud, 2024), DeepL Translate (DeepL) (DeepL, 2024), Google Translate within the Deep Learning Translator (Deep) (Deep Translator, 2020), and OPUS Machine Translation (OPUS) (Tiedemann and Thottingal, 2020), and evaluates the performance of their MT outputs using ConfliBERT (Hu et al., 2022), a domain-specific LLM specialized on political conflict.

This research offers several contributions. The study carefully evaluates the quality of various MT tools using a domain-specific parallel corpus in English, Arabic, and Spanish. Contrary to the ex-

Proceedings of Machine Translation Summit XX Volume 1, pages 315–332, June 23-27, 2025

pectation that LLMs work better when processing native-speaker written/translated texts (NST), experiments in this study indicate that ConfliBERT performs better using MT output. By disentangling sentence-level characteristics, the analysis reveals MT distortions related to nouns, verbs, lemmas, vocabulary complexity, and sentence structure. The study further explores the effects of MT distortions on LLM performance at the sentence level using regression analysis. The results reveal that MT distortions, primarily the vocabulary loss of general and domain-specific rarity, generate simplified text representations that artificially boost model performance, particularly for translated text from Arabic and Spanish to English. Such simplification favors model performance on MT output over NST. These results represent a double-edged sword for researchers using MT tools who may face a trade-off between achieving higher LLM performance at the expense of domain-specific words that may be relevant to their subject of study.

2 Related Works

Pre-trained Language Models (PLM) such as BERT (Devlin et al., 2018a) achieve great results by using continued pre-training on domain-specific data to capture its unique vocabulary, semantics, and language style (Gururangan et al., 2020). Taking advantage of this capability, political scientists use different models to study party manifestos (Mens and Gallego, 2024), voter partisanship (Potter et al., 2024), social movements (Caselli et al., 2021; Hürriyetoğlu et al., 2022; Radford, 2020), dictionary development (Radford, 2021; Osorio et al., 2019), codebook-based classification (Halterman and Keith, 2024) and annotations (Ziems et al., 2024), among other tasks. Similarly, conflict scholars use specialized language to study political violence and cooperation. For that purpose, ConfliBERT (Hu et al., 2022) is a domain-specific model specialized on political conflict that yields superior performance compared to generic LLMs.

Researchers using non-English text generally rely on MT tools to pre-process the original text. However, MTs can heterogeneously distort the data, thus affecting the model performance (Osorio et al., 2024). Assessing MTs from Spanish to English, previous research shows a net summarization of the original text, which reduces the verboseness of the original text and results in an adaptation of the target text to linguistic characteristics of English. This adaptation to English linguistic standards yields high quality-metric results for MT text (Osorio et al., 2024). The summarization effect can further artificially enhance ConfliBERT EN's performance, as the English language generally favors more concise text (Yang et al., 2023).

When evaluating the translation quality from Arabic to English, Osorio et al. (2024) found that MTs artificially extend the source text. This data increase in MTs from Arabic is penalized in English, as metrics show a notable decrease in translation quality relative to the original Arabic text. However, this data increase appears to introduce more linguistic elements that artificially boost ConfliBERT EN's performance on the MT corpus.

The predominant body of research is in favor of languages with abundant resources; thus, more recent studies use translation tools to mitigate the scarcity of training data, including (De Vries et al., 2018), which used Google Cloud (2024) (GT) to translate official transcripts of European Parliament debates written in the official majority of the EU's languages into English. To improve inference in prompting multilingual LLMs, Etxaniz et al. (2023) translated from languages that are comparatively less represented in available LLMs, like Spanish, into English to leverage the fact that English makes up the majority of training data in multilingual LLMs. Other recent studies further compare translation tools (Ibrahim, 2021; Akki and Larouz, 2021; Behr and Braun, 2023) and quality translation metrics (Mathur et al., 2020; Sabtan et al., 2021; He et al., 2021; Lee et al., 2023).

3 Data and Annotations

This research uses the United Nations Parallel Corpus (UNPC) (Ziemski et al., 2016), containing 86,307 official United Nations (UN) Security Council documents translated by professional UN translators. Since the UN operates in six official languages, these translations are considered the Gold Standard Record (GSR). Out of the official UN languages, this study uses NST texts written in English (EN), Spanish (ES), and Arabic (AR). In total, the UNPC contains 11,365,709 fully aligned sentences across languages. This study uses a random sample of 11,326 sentences from UN Security Council documents related to human rights, the protection of civilians, and terrorism. The resulting sample provides a uniquely valuable multilingual parallel corpus in the domain of political conflict and co-

operation. Having the same GSR content across multiple NST sentences within the UNPC provides a *ceteris paribus* leveled field to compare the effects of different MT tools on model performance.

This study uses the UNPC sentences previously annotated by Osorio et al. (2024),[1] which classify the content of the sentences according to PLOVER (Open Event Data Alliance, 2018), an ontology often used in political science to categorize different types of material and verbal interactions based on the cooperative or conflictive conduct of the parties involved. Annotators classified the full sample of sentences according to three tasks. *Relevance* is a binary classification identifying whether a sentence is relevant for political conflict or cooperation or not. *QuadClass* is a multi-class classification task categorizing whether the sentences indicate verbal conflict, verbal cooperation, material conflict, material cooperation, or non-relevant sentences. Finally, *BinQuad* is a binary classification for each QuadClass category identified above, indicating whether the sentence can be categorized as the respective PLOVER category or not, thereby representing one of the other three categories.

The annotations have the following distributions[2]. In the Relevance binary task, coders identified 52% sentences as not relevant and the rest 48% as relevant. For the multi-class QuadClass task, coders identified 14% sentences as Material Conflict, 13% as Material Cooperation, 8% as Verbal Conflict, 11% as Verbal Cooperation, and 53% as not relevant. Finally, the BinQuad binary task of QuadClass categories produced the following distribution for Material Conflict (yes 14%, no 86%), Material Cooperation (yes 13%, no 87%), Verbal Conflict (yes 8%, no 92%), and Verbal Cooperation (yes 11%, no 89%). All experiments used balanced datasets, with the number of randomly selected sentences capped to match the smallest category size in each task.

4 Translation Quality Assessment

Using UNPC text in English (EN), Spanish (ES), and Arabic (AR), we conduct a series of MTs using different tools. Our analysis uses bidirectional MT to convert the entire sample of Spanish and Arabic texts into English (ES to EN, AR to EN) and vice versa (EN to ES, EN to AR). To conduct the translations, we use four commonly used MT tools: Google API Translate (GT) (Google Cloud, 2024), DeepL Translate (DeepL) (DeepL, 2024), Deep Learning Translator (Deep) (Deep Translator, 2020), and OPUS Machine Translation (OPUS) (Tiedemann and Thottingal, 2020).[3] Google translate employs subword tokenizers optimized on extensive multilingual corpora (Kudo and Richardson, 2018). This approach addresses out-of-vocabulary challenges by merging frequent character sequences into subwords and transforming tokenized subwords into dense embeddings within the Google-managed Transformer architecture. Positional encodings enable the self-attention mechanism to align and predict tokens in the encoder-decoder pipeline. These vectors are continually refined through large-scale training on vast datasets, a process referred to as dynamic tuning (Google Cloud, 2024; Vaswani et al., 2017). DeepL provides free and subscription-based translating services between a variety of languages via its website or an API (DeepL, 2024). Deep is a lightweight Python package that invokes the public Google Translate service. It accesses a standard and universal shared model that lacks dynamic tuning capabilities. This causes lower accuracy or a failure to accurately capture complex and domain-specific words (Deep Translator, 2020; Google Cloud, 2024). Google Translate was selected via the deep translation package to establish a baseline comparison between the paid and free versions of the most used MT tool in the literature (Wu et al., 2016). OPUS, a Hugging Face Transformers library, presents a suite of state-of-the-art pretrained translation models (Tiedemann and Thottingal, 2020). In particular, its Helsinki-NLP/opus-mt-ar-en and Helsinki-NLP/opus-mt-es-en models are specifically trained to translate from Arabic to English and from Spanish to English, respectively (OPUS, 2016).

Building on (Han et al., 2022), we use four quality assessment metrics: SacreBLEU (Post, 2018), METEOR (Banerjee and Lavie, 2005), BERTScore (Devlin et al., 2018b), and COMET (Crosslingual Optimized Metric for Evaluation of Translation), using the wmt20-comet-da model (Rei et al., 2020). SacreBLEU, and METEOR are lexical-based metrics measuring the similarity between NST and MT text using mathematical or heuristic methods, COMET is a neural-based metric (Rei et al., 2020),

[1] See appendix D for details on the annotation process.
[2] Details in Appendix E
[3] See Appendix F on MT tools development, training data relevance, and Appendix G for quality metric evaluation.

whereas BERTScore uses an embedding-based metric that relies on deep learning methods (Lee et al., 2023). Quality scores range from 0 to 1, with high values indicating greater NST-MT similarity (Chatzikoumi, 2020; Zhang et al., 2020). The metrics can be ranked according to their degree of flexibility. SacreBLEU employs the Moses tokenizer, an advanced preprocessing tool that facilitates score comparability and was first created for the Moses statistical machine translation system (Post, 2018). The Moses tokenizer uses heuristics and rules unique to a given language to normalize text and to handle punctuation or special characters. METEOR is more flexible and calculates the similarity of word alignments. COMET is a state-of-the-art neural-based MT evaluation metric. BERTScore is the most flexible metric as it considers contextual correctness and synonyms.

Using each UNPC NST text as a reference, Figure 1 presents the quality scores from the different metrics applied to each MT output. Results show that different tools generate varying degrees of quality across languages. While for AR to EN and ES to EN MTs, SacreBLEU, METEOR, and BERTScore indicate that DeepL provides the best quality output, COMET considers OPUS to be the most accurate MT tool for these language combinations. For EN to ES, OPUS yields the best MT quality based on all metrics, while for EN to AR, COMET disagrees with all other metrics and considers DeepL the best-performing MT tool. While these metrics offer a first assessment of the MT quality, they do not permit an in-depth understanding of MT-induced translationese effects on the source text or assess more subtle changes in meaning and nuance. Consequently, these metrics do not fully capture whether MT-induced changes influence LLM performance. The following section evaluates LLM performance across MT texts to see if the results align with quality assessment suggestions.

5 Model Performance Across MT Tools

Following the quality assessment, we test the effect of MTs on LLM performance. To do so, we use ConfliBERT (Hu et al., 2022), a domain-specific pre-trained language model specifically designed to analyze political texts, to evaluate the UNPC NST and MT texts for three classification tasks: Relevant (binary) classification, QuadClass (multiclass) classification, and BinQuad (binary) classification of each QuadClass category. For each task, the fine-tuning uses three versions of the ConfliBERT family, namely ConfliBERT Arabic (Alsarra et al., 2023), ConfliBERT Spanish (Yang et al., 2023), and ConfliBERT English (Hu et al., 2022), with cased and uncased variations, resulting in a total of 6 different models. All models use balanced datasets for each task. By keeping the UNPC content and the use of ConfliBERT constant, we analyze variations in performance derived from different MT tools, including Deep, DeepL, GT, OPUS, and the NST texts. First, we split the data into training, testing, and developing using 70-15-15 rule. Second, for each model, we perform the evaluation using 10 seeds and 5 epochs. Finally, we run a total of 114 fine-tuning tasks on those models and their corresponding datasets. We used a HPC system with a single A100 GPU 20GB and a single V100 GPU 32GB, and a learning rate of 4e-05, with a training batch size of 8 and a maximum sequence length of 512 for both binary and multiclass classifications. Figures 2-4 present the F1 scores highlighting the top-performing models in red. Overall, results show that processing MT text yields better results than analyzing NST text. This is puzzling since domain-specific models would be expected to perform better with NST texts.

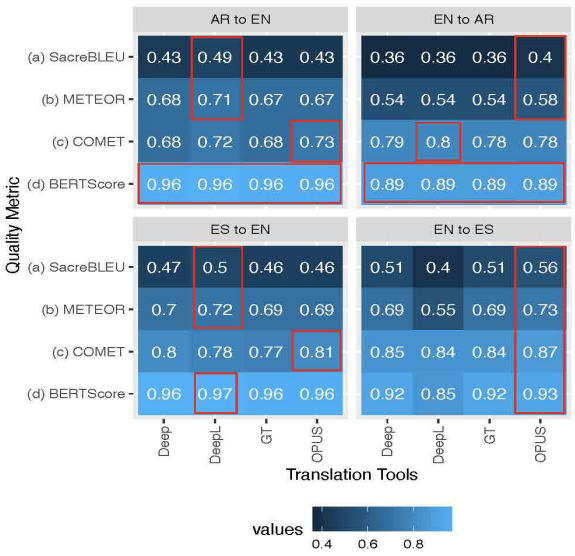


Figure 1: Quality Assessment Metrics

5.1 Relevant Binary Classification

Figure 2 reports the F1 performance of the ConfliBERT models for the relevant binary task on the NST and MT texts across languages. Red squares indicate best models with p-values at $p<0.01$ or lower. Overall, the results show high performance

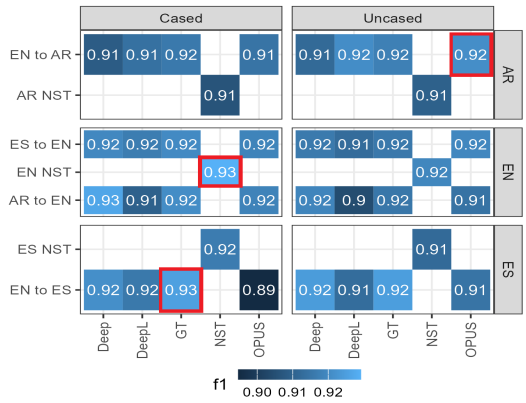


Figure 2: Binary Relevance Classification

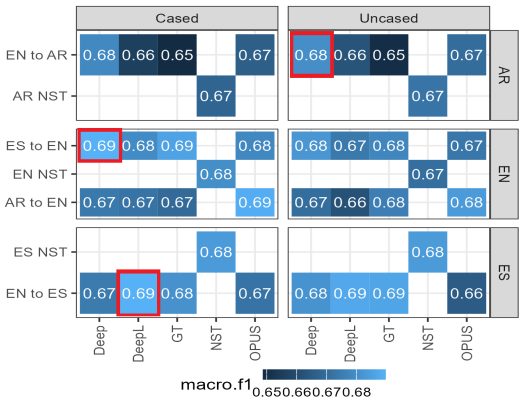


Figure 3: QuadClass Classification

levels and little variations across MT outputs. Most importantly, the results show that MT texts yield marginally better results in Arabic and Spanish than models processing the NST texts. Yet, these results do not hold for English, where models using NST text outperform those analyzing MT text. For Arabic, ConfliBERT uncased performs best (F1 0.921) on the EN to AR corpus translated with OPUS. This result is statistically significantly better than NST text in Arabic (F1 0.91). For English, ConfliBERT cased on English NST text shows the highest performance (F1 0.929) and performs statistically significantly better than AR to EN Deep (F1 0.927). Finally, the results for Spanish indicate that ConfliBERT cased performs best using the EN to ES GT translation (F1 0.925) and significantly better than NST text in Spanish (F1 0.917).

5.2 QuadClass Multi-Class Classification

Figure 3 presents the results of the QuadClass classification task. Overall, the results of the QuadClass classification report lower performance than the Relevant binary task. This is understandable as a five-categories multi-class classification is more difficult than a dichotomous task, and the latter has more training examples than the former. In general, these results also indicate that analyzing MT text performs marginally better than processing NST texts across languages. For Arabic, ConfliBERT uncased reports the highest F1 (0.680) for the QuadClass on the EN to AR Deep translated text. This result is statistically significantly better than the NST text Arabic model (F1 0.672). For English, ConfliBERT cased processing the ES to EN Deep translation generates the best QuadClass performance (F1 0.69), while the NST model in English (F1 0.68) has a statistically significantly lower performance. Finally, the results for Spanish show that ConfliBERT cased performs best using EN to ES DeepL translation (F1 0.69). This result is barely better than Spanish NST text (F1 0.684).

5.3 BinQuad Binary Classification

Figure 4 shows binary classification results for Material Conflict (panel 4.a), Material Cooperation (panel 4.b), Verbal Conflict (panel 4.c), and Verbal Cooperation (panel 4.d). The analysis shows heterogeneous results. For specific QuadClass instances in Arabic, MT generally performs better than Arabic NST text. However, the results for NST text versus MT text in English and Spanish are mixed.

Panel 4.a shows the Material Conflict scores. In general, the results indicate that MT text performs better than NST text in Arabic and English, but the Spanish models show a comparable performance in NST and MT texts. For Arabic, ConfliBERT uncased has the best performance (F1 0.885) when processing the Deep EN to AR translation. In contrast, NST text Arabic has a statistically lower performance (F1 0.863). For English, ConfliBERT cased performs the best (F1 0.908) using the ES to EN Deep text. This result is statistically significantly better than using English NST text (F1 0.890). For Spanish, ConfliBERT cased reports the best performance with the EN to ES OPUS text (F1 0.876). However, this result is not statistically different from the NST text in Spanish (F1 0.876).

Material Cooperation results (panel 4.b) indicate that MT works as well as NST Arabic and English texts and sometimes works better than Spanish NST text. For Arabic, ConfliBERT cased with NST Arabic text yields the best performance (F1 0.819). Yet, it is not different from ConfliBERT uncased with EN to AR OPUS translation (F1 0.818). For English, the top performing model is ConfliBERT

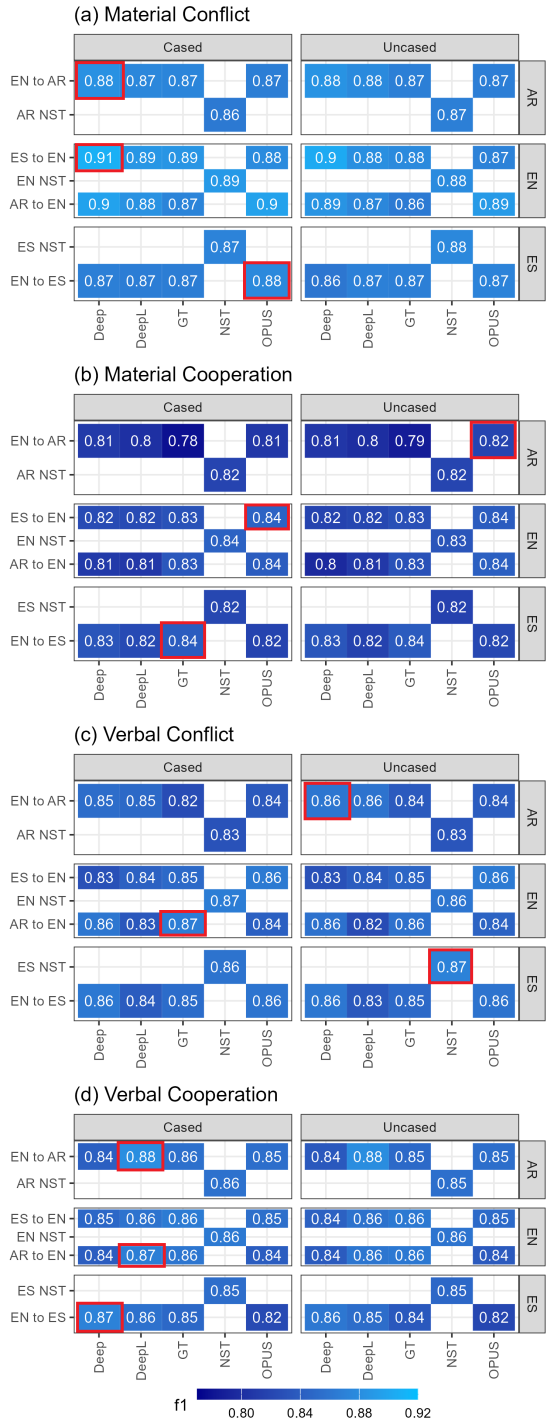


Figure 4: Binary QuadClass Classification

cased using the ES to EN OPUS translation (F1 0.844). However, this result is not statistically superior to English NST text (F1 0.843). Finally, the results for Spanish show that the EN to ES GT text yields the best results with ConfliBERT cased (F1 0.838), while the Spanish NST text model has lower performance (F1 0.82).

Panel 4.c reports the Verbal Conflict F1 scores. In general, the results show that MT texts perform better in Arabic, but these findings do not hold for English and Spanish. The results for Arabic indicate that ConfliBERT uncased has the best performance with EN to AR Deep text (F1 0.86). This score is better ($p<0.001$) than the Arabic NST text performance (F1 0.833). For English, the AR to EN GT translation using ConfliBERT cased has the best result (F1 0.867). However, this score is not different ($p=0.805$) from the English NST text model (F1 0.866). For Spanish, ConfliBERT uncased works the best when using Spanish NST text (F1 0.87). However, this result is not different ($p=0.197$) from its closest competitor, the EN to ES Deep text with ConfliBERT cased (F1 0.862).

Finally, Verbal Cooperation results in panel 4.d indicate that MT texts yield better results than NST texts in Arabic and Spanish, but models using sentences in English perform as well as those using MT texts. For Arabic, ConfliBERT uncased processing EN to AR DeepL translations has the best performance (F1 0.88). This result is statistically superior to the Arabic native model (F1 0.856). For English, the top performing model is ConfliBERT cased processing AR to EN DeepL translation (F1 0.867). Although this model performs better than the English NST text model (F1 0.863), the difference is not statistically significant. Finally, the results for Spanish indicate that processing the EN to ES Deep translation with ConfliBERT cased yields the best performance (F1 0.87). In contrast, the Spanish NST text model reports a lower performance (F 0.853) at statistically significant levels.

Overall, this section shows that LLM performance does not necessarily align with the MT quality suggestions. The following sections try to identify the determinants of model performance.

6 Corpus Rarity and Vocabulary Loss

To disentangle the characteristics of MT outputs that yield marginally superior ConfliBERT performance compared to processing NST texts, this section analyzes MT-induced distortions to the original corpora. First, we measure the total vocabulary size after preprocessing using spaCy's `en_core_web_trf` transformer pipeline for English (Honnibal et al., 2020), spaCy's `es_dep_news_trf` transformer pipeline for Spanish (Honnibal et al., 2020), and the Farasa segmenter (Al-shaibani, 2021) for Arabic. The vocabulary size for each language represents the total number of unique words included in the MT corpora. Figure 5 shows the vocabulary sizes. We

find that the MT corpora consistently have a lower vocabulary size than the respective NST corpus. This finding aligns with the characteristics of translationese, where translated text tends to show reduced lexical diversity compared to original text (Riley et al., 2020). Therefore, there may be a convergence of MT on similar phrasing, reducing the need to learn diverse patterns as in the native text.

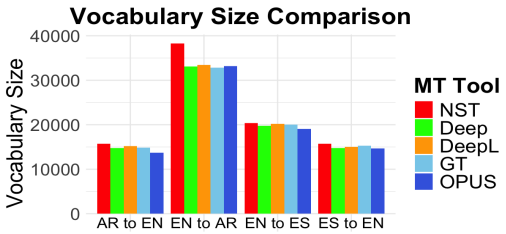


Figure 5: Native-MT Vocabulary Size Comparison

To further explore this reduction effect, we measure lexical rarity per sentence. Following (Proisl, 2022), we define lexical rarity as the proportion of tokens in a text that does not appear in the 5,000 most common tokens for a domain. We consider two types of rarity: general and domain. General rarity relies on the 5,000 most common tokens for a language, regardless of subject. Domain rarity uses the 5,000 most common tokens in the sentences from the UNPC to measure rarity as it relates to a political corpus. We use rarity as a proxy for lexical complexity and consider it the prime indicator for the reduction of text complexity and loss of context in the MT texts. A reduction in rarity for MT text represents a decline in the number of unique tokens compared to the NST text. Consequently, a reduction in the mean rarity of the MT corpus represents a loss of language complexity compared to the NST corpus. The loss of rarity may be particularly relevant for domain-specific researchers where key terms or technical words may carry particular substantive value. In addition to rarity, we measure the number of unique lemmas, nouns, and verbs in each sentence as additional measures of linguistic features (see Appendix H).

Figure 6 shows the mean general and domain rarity scores. Using a pairwise Wilcox test from the `stats` R package (R Core Team, 2023), we compare the MT text to the NST corpora in terms of rarity and find that the English and Arabic MTs all have statistically significantly lower general and domain rarity scores than the respective NST corpus. Spanish, however, does not display the same effect. In contrast, MTs using Deep, DeepL, and GT all have statistically significant higher general rarity scores than the Spanish NST corpus. However, regarding domain rarity, the difference between the Spanish NST text and the MT output using Deep, DeepL, and GT is not significant. OPUS translations are not significantly different from Spanish NST text in general rarity, but show a statistically significant reduction in domain rarity.

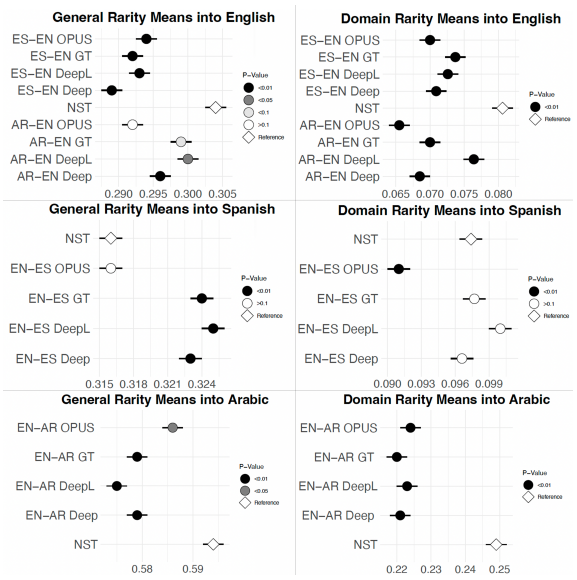


Figure 6: General and Domain Rarity Means

These results and the reduced vocabulary size show that MTs into English and Arabic generate in a significant loss of rare tokens. While this loss may simplify the text and facilitate model classification, these translationese-induced simplifications may lead to the loss of critical context where specific words and their substantive meaning are essential.

The significant increase in rarity for some MT tools into Spanish is likely due to a bias toward brevity using more complex words. While native speakers may opt for longer but simpler phrasing in NST text, the MT text may result in a brief but vocabulary-heavy phrasing. This may be due to the training data for translation into Spanish favoring these characteristics. The reduction in text complexity further indicates that there could be overfitting to MT text in fine-tuning. Models trained on the MT text, which has lower text complexity, may not perform well when tasked with classifying NST text or even text from another MT tool that introduces higher or different types of complexity. This finding, consequently, warrants additional consideration when fine-tuning using MT text.

These findings resonate with major challenges in Neural Machine Translation (NMT). First, NMT

systems perform poorly in specialized domains for which the system has not been trained for. Second, NMT systems are weak at translating low-frequency (rare) words, especially in cases where there are many inflections (as in verb conjugations in Spanish). Third, NMT systems struggle with long sentences, which are disproportionately underrepresented in the UNPC (Koehn and Knowles, 2017). Finally, Vanmassenhove et al. (2019) similarly find that MT systems fail to reach the diversity of phrasing and vocabulary of natural human language. Therefore, there is a loss of information, context, and the unique voice of the speaker/writer of the source text in this process. While event classification may not suffer from this loss, other tasks, such as Named Entity Recognition (NER), may experience poorer performance using MT.

7 Dependency Distance

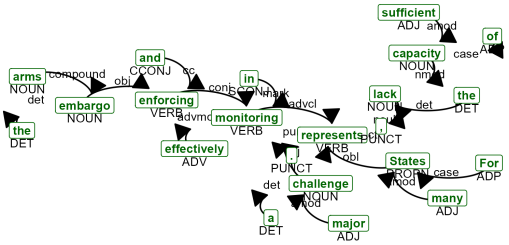
To explore the MT effects on model performance, the study also analyzes the dependency distance mean (DDM) of each sentence across languages and the dependency distance mean difference (DDM^d) of each MT output to their corresponding NST text. DDM is the average syntactical distance between the root of a sentence to other parts of speech and is generally regarded as an indicator of sentence complexity (Liu et al., 2017, 2022). A high DDM refers to highly complex sentences. Relatedly, DDM^d is interpreted here as the distortion caused by the MT tool when compared to its target NST text, such that a negative DDM^d indicates syntactical simplification and a positive DDM^d shows increasing syntactical complexity by the MT tool.

To get the DDM, we use `textdescriptives`, `spacy`, `spacy_transformers`, and libraries with `en_core_web_sm`, `bert-large-arabertv02`, and `es_core_news_sm` models for English, Arabic, and Spanish, respectively. For example, Table 1 presents the same English NST sentence in comparison to its English MT using DeepL from Arabic and Spanish texts. As Table 1 shows, small differences in the MT output are consequential for the dependency tree, the DDM, and DDM^d of the MTs into English. See Appendix I for details.

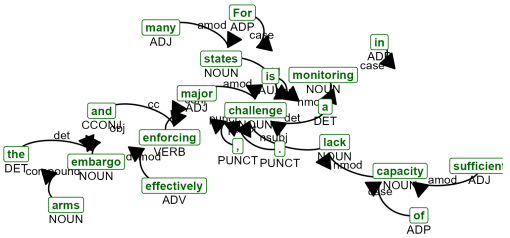
8 Sentence-Level Prediction Confidence

To better understand ConfliBERT's performance across NST and MT texts, we analyze the effects of different sentence-level characteristics on model performance. We first estimate the degree of con-

(a) EN NST: "For many States, the lack of sufficient capacity represents a major challenge in effectively monitoring and enforcing the arms embargo." DDM = 2.63.



(b) AR to EN using DeepL: "For many states, lack of sufficient capacity is a major challenge in monitoring and effectively enforcing the arms embargo." DDM = 3.09, DDM^d compared to EN NST = 0.46.



(c) ES to EN using DeepL: "The lack of sufficient capacity is a major challenge for many States in the effective monitoring and enforcement of arms embargoes." DDM = 2.45, DDM^d compared to EN NST = -0.18.

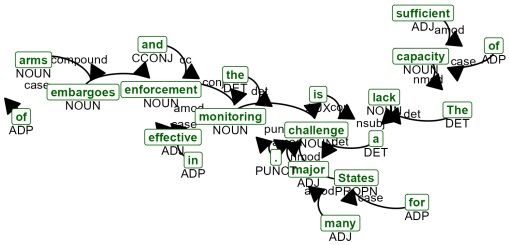


Table 1: Dependency Distance Example

fidence to which ConfliBERT correctly classifies each sentence. Then, we use regression analysis to explain the levels of prediction confidence based on various sentence-level characteristics.

To calculate prediction confidence at the sentence level, we use the ConfliBERT-uncased model in the Binary Relevant classification task. The methodology generates label predictions and confidence scores by applying the softmax function to its output logits (Devlin et al., 2018b). The methodology processes the logits of each sentence through softmax, converting them into probabilities ranging from (0,1), thus indicating the model's confidence in correctly assigning the chunk to a specific class. In this way, the prediction reflects the probability of ConfliBERT's correct classification.

For sentences longer than 512 tokens, we apply a chunking strategy, splitting them into segments

of 512 tokens each (Pappagari et al., 2019; Park et al., 2022), and independently classify and generate a predicted label and confidence score for each segment. To ensure an accurate sentence-level prediction, we apply majority voting to determine the final label and average the confidence scores across all chunks in a sentence. This method ensures that the final confidence prediction reflects the model's certainty across the entire sentence, allowing us to handle longer texts without losing context or compromising accuracy. Averaging the confidence scores provides a robust measure of the model's overall confidence.

9 Explaining Model Performance

To explore the determinants of model performance, we further analyze the sentence-level prediction confidence for the binary classification task using a linear regression model as indicated in equation 1.

$$y_i = \alpha + \beta_1 V_i + \beta_2 N_i + \beta_3 L_i + \beta_4 R_i^g + \beta_5 R_i^d + \beta_6 DDM_i + \beta_7 DDM_i^d + \epsilon_i \quad (1)$$

where y_i is the predicted confidence of ConfliBERT correctly identifying the binary classification for sentence i. The independent variables refer to sentence characteristics that could affect model performance, including the number of verbs (V_i), the noun count (N_i), unique lemmas (L_i), general rarity (R_i^g), domain rarity (R_i^d), the dependency distance mean (DDM_i), and the DDM difference (DDM_i^d) caused by MT, the latter is only included in MT texts. α and ϵ represent the intercept and the errors, respectively. To facilitate the comparison of coefficients, we standardize V_i, N_i, L_i, and DDM_i to range from 0,1 for the count measures, and a [-1,1] range for DDM_i^d. Using equation 1, we regress these sentence-level characteristics to explain ConfliBERT's performance for the binary classification task across NST and MT outputs. Appendix J reports the regression results.

Following Ward et al. (2010) and Brandt et al. (2022), we evaluate the contribution of each variable on the probability of correct classification by comparing the contribution of each sentence-level characteristic to the regression Root Mean Standard Error (RMSE) using stepwise elimination. RMSE is the standard deviation of the residuals away from the regression line. A low RMSE indicates that the observations closely revolve around the regression line, which suggests a good model fit. The stepwise elimination approach consists of first running the full regression and calculating the RMSE, then dropping one variable at a time from equation 1 and comparing the change in the RMSE from each subsequent model. A large RMSE increase after eliminating a certain variable indicates a greater model fit loss, suggesting that this variable largely contributes to the model performance. Since each regression has its own RMSE (see Appendix K), we favor the comparability of results by calculating the Model Fit Loss as a percentage using as baseline the full model's RMSE. This provides a standardized measure for cross-model comparison in which lower Model Fit Loss values indicate worse model performance after each variable elimination.

Figure 7 presents the Model Fit Loss by stepwise elimination across native and MT texts. The baseline in each panel is the full model RMSE from equation 1. The Model Fit gradually decreases after subsequently dropping each independent variable in each elimination step; the magnitude of the drop indicates the contribution of each eliminated variable. In general, Figure 7 shows that all models experience substantial performance loss after eliminating the general and domain rarity variables. This shows that general rarity and domain-specific rarity have considerable leverage in explaining ConfliBERT performance for binary classification. Text translated using OPUS seems particularly sensitive to the contributions of general and domain rarity in any translation direction.

These results offer an important finding suggesting that highly specialized words are crucial for explaining model performance. As MT tools generally reduce the vocabulary richness (see Figure 5) and decrease the number of specialized or domain-specific terms in the MT text (see Figure 6), the resulting translation output is a simplified representation of the native text containing sentences with fewer tokens and simpler words. This MT text generally makes it easier for LLMs to process. However, this performance gain comes at the cost of lower vocabulary richness in key terms.

10 Discussion and Conclusion

Table 2 presents a general summary of the main results. Based on these findings, we derive the following main conclusions: First, MT quality assessment scores provide limited insight about which MT tool performs best across classification tasks. Most quality scores recommend DeepL and OPUS as the best tool for Arabic and English, and OPUS

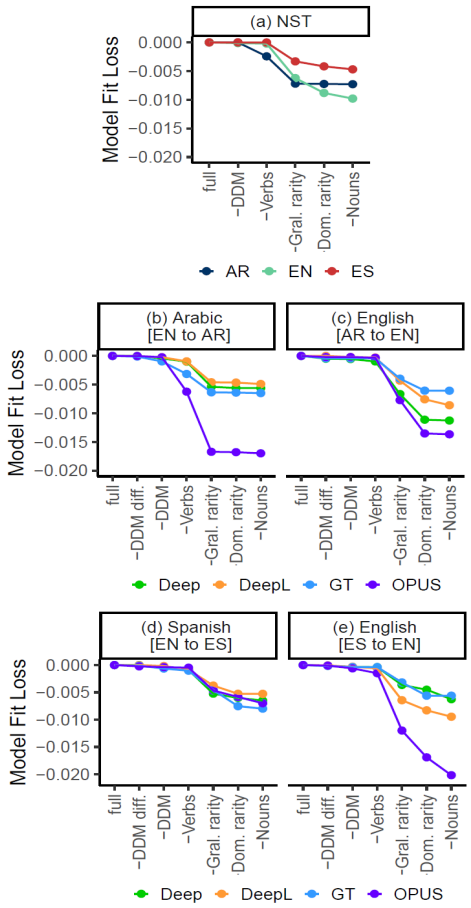


Figure 7: Model Fit Loss by Stepwise Elimination

for Spanish. However, these tools rarely outperform other MT texts across classification tasks.

Second, the sentence-level analysis reveals that all MT tools induce a reduction in vocabulary complexity. In addition, Arabic and English translations suffer from a reduction in both general and domain-specific lexical rarity. This suggests an important simplification of key terms that may be of particular relevance to domain experts. However, we detect an increase in general rarity in Spanish, and no general changes in domain rarity.

Third, although LLMs are expected to perform better with native documents than with MT texts, results across downstream tasks indicate that LLMs generally perform better with MT texts than with native corpora. Yet, no single MT tool consistently reports the top performance across languages.

Finally, using regression analysis and a stepwise deletion approach to assess the contributions of different sentence-level characteristics on model performance, the analysis indicates that highly specialized words—represented by general and domain-specific rarity—have the most leverage in explaining model performance for binary tasks.

Based on a specific application in the field of political science, this study suggests an important trade-off for the use of MT tools that could be extended to other domains. On the one hand, results indicate that MT tools may substantially reduce the time and effort for human analysis to process large volumes of text, and such MT texts tend to yield better results when using specialized LLMs for a variety of tasks. In simple terms, it seems that machines talking to machines tend to generate better results. On the other hand, the use of MT tools tends to produce translationese outputs that reduce vocabulary richness, particularly for rare terms that may be of high substantive value to domain experts. For human translators operating in highly technical fields, such vocabulary loss may prove unacceptable despite the artificially superior machine-to-machine performance.

Finding	Arabic	English	Spanish
Best MT Quality	OPUS	DeepL	OPUS
MT Voc. size	Decrease	Decrease	Decrease
Gral. rarity	Decrease	Decrease	Increase
Domain rarity	Decrease	Decrease	Not signif.
Best Binary	OPUS	Native	GT
Best QuadClass	Deep	Deep	DeepL
Best Mat. Conf.	Deep	Deep	OPUS
Best Mat. Coop.	OPUS	OPUS	GT
Best Verb. Conf.	Deep	GT	Native
Best Verb. Coop.	DeepL	DeepL	Deep
Main performance contributors	R_i^g	R_i^g and R_i^d	R_i^g and R_i^d

Table 2: Summary of Results

11 Limitations

The study has several limitations. First, this analysis is circumscribed to the political domain. Therefore, results may not be generalizable to other domains. Second, the conclusions derived from the regression analysis are based on a relatively simple binary classification task. Future research should evaluate if these findings hold in more sophisticated downstream tasks such as multi-class and multi-label classification, or named entity recognition. Third, MT tools were trained on either the UNPC itself (OPUS) or similar multilingual UN text (GT, Deep), or can be expected to have been trained on it (DeepL). MT tools can, therefore, be expected to achieve a higher translation accuracy due to their familiarity with UN text. Furthermore, some MT tools are not free, limiting their acceptability. Additionally, the study does not include other MT tools such as ChatGPT (OpenAI, 2022) or Gemini (Gimine, 2023). However, the selection of MT tools focuses on the most used tools in the chosen languages. Fourth, fine-tuning ConfliBERT on multiple languages with large datasets consumes significant time and computational resources.

References

Lars Ahrenberg. 2017. Comparing machine translation and human translation: A case study. In *Proceedings of the Workshop Human-Informed Translation and Interpreting Technology*, pages 21–28, Varna, Bulgaria. Association for Computational Linguistics.

Fouad Akki and Mohammed Larouz. 2021. A comparative study of english-arabic-english translation constraints among efl students. *International Journal of Linguistics and Translation Studies*, 2(3):33–45.

Maged Saeed Al-shaibani. 2021. Magedsaeed/farasapy: A python implementation of farasa toolkit. https://github.com/MagedSaeed/farasapy. (Accessed on 04/16/2025).

Sultan Alsarra, Luay Abdeljaber, Wooseong Yang, Niamat Zawad, Latifur Khan, Patrick Brandt, Javier Osorio, and Vito D'Orazio. 2023. ConfliBERT-Arabic: A pre-trained Arabic language model for politics, conflicts and violence. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 98–108, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

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.

Osvaldas Bartaškevičius. 2024. *Light machine translation post-editing effort and quality evaluation of news texts translated from English to Lithuanian*. Ph.D. thesis, Kauno technologijos universitetas.

Dorothée Behr and Michael Braun. 2023. How does back translation fare against team translation? an experimental case study in the language combination english–german. *Journal of survey statistics and methodology*, 11(2):285–315.

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, and James Starz. 2018. ICEWS Weekly Event Data.

Patrick T. Brandt, Vito D'Orazio, Latifur Khan, Yi-Fan Li, Javier Osorio, and Marcus Sianan. 2022. Conflict forecasting with event data and spatio-temporal graph convolutional networks. *International Interactions*, 48(4):800–822. Accessed: 2024-08-27.

Giulia Cambedda, Giorgio Maria Di Nunzio, and Viviana Nosilia. 2021. A study on automatic machine translation tools: A comparative error analysis between deepl and yandex for russian-italian medical translation. *Umanistica Digitale*, (10):139–163.

Tommaso Caselli, Osman Mutlu, Angelo Basile, and Ali Hürriyetoğlu. 2021. Protest-er: Retraining bert for protest event extraction. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 12–19, Online. Association for Computational Linguistics. Accessed: 2023-07-10.

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Erik De Vries, Martijn Schoonvelde, and Gijs Schumacher. 2018. No longer lost in translation: Evidence that google translate works for comparative bag-of-words text applications. *Political Analysis*, 26(4):417–430.

Deep Translator. 2020. deep-translator: A flexible free and unlimited python tool to translate between different languages in a simple way using multiple translators. https://github.com/nidhaloff/deep-translator. (Accessed on 04/16/2025).

DeepL. 2024. DeepL Translator. https://www.deepl.com/translator. (Accessed on 04/16/2025).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english? *Preprint*, arXiv:2308.01223.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. In L. Wollin and H. Lindquist, editors, *Translation studies in Scandinavia: Proceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II*, number 75 in Lund Studies in English, pages 88–95. CWK Gleerup, Lund.

Gimine. 2023. Gimine: Open-source data mining platform. https://gimine.com. Accessed: 2024-09-12.

Google Cloud. 2024. Google cloud translation api. Accessed: 2025-04-16.

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*.

Andrew Halterman and Katherine A. Keith. 2024. Codebook llms: Adapting political science codebooks for llm use and adapting llms to follow codebooks. *Preprint*, arXiv:2407.10747.

Andrew Halterman, Philip A. Schrodt, Andreas Beger, Benjamin E. Bagozzi, and Grace I. Scarborough. 2023. Creating custom event data without dictionaries: A bag-of-tricks. *arXiv preprint arXiv:2304.01331*.

Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff, and Goran Nenadic. 2022. Examining large pre-trained language models for machine translation: What you don't know about it. *Preprint*, arXiv:2209.07417.

Pinjia He, Clara Meister, and Zhendong Su. 2021. Testing machine translation via referential transparency. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, pages 410–422.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. Accessed: 2025-04-16.

Yibo Hu, MohammadSaleh Hosseini, Erick Skorupa Parolin, Javier Osorio, Latifur Khan, Patrick Brandt, and Vito D'Orazio. 2022. Conflibert: A pre-trained language model for political conflict and violence. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5469–5482.

Yan Huang and Wei Liu. 2024. Evaluating the Translation Performance of Large Language Models Based on Euas-20. *arXiv preprint*. ArXiv:2408.03119 [cs].

Ali Hürriyetoğlu, Osman Mutlu, Fırat Duruşan, Onur Uca, Alaeddin Gürel, Benjamin J. Radford, Yaoyao Dai, Hansi Hettiarachchi, Niklas Stoehr, Tadashi Nomoto, Milena Slavcheva, Francielle Vargas, Aaqib Javid, Fatih Beyhan, and Erdem Yörük. 2022. Extended Multilingual Protest News Detection - Shared Task 1, CASE 2021 and 2022. In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 223–228, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dalia Ibrahim. 2021. Google translation and the question of ideology in political news headlines. *SAHIFATUL-ALSUN*, 37(37):57–80.

Navroz Kaur Kahlon and Williamjeet Singh. 2023. Machine translation from text to sign language: a systematic review. *Universal Access in the Information Society*, 22(1):1–35.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. ArXiv preprint arXiv:1706.03872.

T. Kudo and J. Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proc. 2018 Conf. Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*, pages 66–71.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the Carbon Emissions of Machine Learning. *arXiv preprint*. ArXiv:1910.09700 [cs].

Sangmin-Michelle Lee. 2023. The effectiveness of machine translation in foreign language education: a systematic review and meta-analysis. *Computer Assisted Language Learning*, 36(1-2):103–125.

Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heuiseok Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics*, 11(4):1006.

Hauke Licht, Ronja Sczepanski, Moritz Laurer, and Ayjeren Bekmuratovna. 2024. No more cost in translation: Validating open-source machine translation for quantitative text analysis. Discussion Paper.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21:171–193.

Shanshan Liu and Wenxiao Zhu. 2023. An analysis of the evaluation of the translation quality of neural machine translation application systems. *Applied Artificial Intelligence*, 37(1):2214460.

Xueying Liu, Haoran Zhu, and Lei Lei. 2022. Dependency distance minimization: A diachronic exploration of the effects of sentence length and dependency types. *Humanities and Social Sciences Communications*, 9(1):1–9.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *CoRR*, abs/2006.07264.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics. *Preprint*, arXiv:2006.06264.

Gaël Le Mens and Aina Gallego. 2024. Positioning political texts with large language models by asking and averaging. ArXiv preprint arXiv:2311.16639.

Open Event Data Alliance. 2018. Political language ontology for verifiable event records. `https://github.com/openeventdata/PLOVER`. Accessed: 2022-10-01.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`. Accessed: 2024-09-12.

OPUS. 2016. An overview of the OPUS collection. `https://opus.nlpl.eu/`. Accessed: 2025-04-05.

Javier Osorio, Sultan Alsarra, Amber Converse, Afraa Alshammari, Dagmar Heintze, Naif Alatrush, Latifur Khan, Patrick T. Brandt, Vito D'Orazio, Niamat Zawad, and Mahrusa Billah. 2024. Keep it local:

Comparing domain-specific llms in native and machine translated text using parallel corpora on political conflict. In *The 2nd International Conference on Foundation and Large Language Models*.

Javier Osorio, Viveca Pavon, Sayeed Salam, Jennifer Holmes, Patrick T. Brandt, and Latifur Khan. 2019. Translating CAMEO verbs for automated coding of event data. *International Interactions*, 45(6):1049–1064.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844.

Hyunji Hayley Park, Yogarshi Vyas, and Kashif Shah. 2022. Efficient classification of long documents using transformers. *Preprint*, arXiv:2203.11258.

Matt Post. 2018. A call for clarity in reporting bleu scores. ArXiv preprint arXiv:1804.08771.

Yujin Potter, Shiyang Lai, Junsol Kim, James Evans, and Dawn Song. 2024. Hidden Persuaders: LLMs' Political Leaning and Their Influence on Voters. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4244–4275, Miami, Florida, USA. Association for Computational Linguistics.

Thomas Proisl. 2022. textcomplexity: Linguistic and stylistic complexity. Accessed: 2025-04-16.

R Core Team. 2023. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Accessed: 2025-04-16.

Benjamin J Radford. 2020. Multitask Models for Supervised Protests Detection in Texts. *arXiv preprint arXiv:2005.02954*.

Benjamin J. Radford. 2021. Automated dictionary generation for political eventcoding. *Political Science Research and Methods*, 9(1):157–171.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

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.

Yasser Sabtan, Mohamed Hussein, Hamza Ethelb, and Abdulfattah Omar. 2021. An evaluation of the accuracy of the machine translation systems of social media language. *International Journal of Advanced Computer Science and Applications*, 12.

Nick Schäferhoff. 2024. The history of google translate (2004–today): A detailed analysis. Accessed: 2025-04-16.

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.

United States Environmental Protection Agency. 2015. Greenhouse gas equivalencies calculator. Accessed: 2025-04-16.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Michael D Ward, Brian D Greenhill, and Kristin M Bakke. 2010. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):363–375. Publisher: SAGE Publications Ltd.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, and others. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wooseong Yang, Sultan Alsarra, Luay Abdeljaber, Niamat Zawad, Zeinab Delaram, Javier Osorio, Latifur Khan, Patrick T. Brandt, and Vito D'Orazio. 2023. Conflibert-spanish: A pre-trained spanish language model for political conflict and violence. In *2023 7th IEEE Congress on Information Science and Technology (CiSt)*, pages 287–292.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *Preprint*, arXiv:1904.09675.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational linguistics - Association for Computational Linguistics*, 50(1):237–291.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Replication Files

The data and replication files are available in GitHub at https://github.com/javierosorio/devil_in_the_details_mtsummit25.

B Ethical Considerations

This research utilizes United Nations Parallel Corpus as a source of information but does not involve human research subjects. By evaluating MT tools on political conflict and cooperation, we aim to help low-resource languages expand their high-quality data on such scarce contents (Magueresse et al., 2020). Creating a gold standard content of UN data aligned per sentence across multiple native languages sets up the foundation for researchers to use as labeled data sets for the specified languages of English, Arabic, and Spanish. And even extrapolate the work into the other three UN official languages, such as French, Chinese, and Russian, which are also considered lower-resource languages when compared to English.

C Sustainability Statement

Following Lacoste et al. (2019), this section presents the estimated energy cost and its corresponding carbon impact statement. The experiments reported in this study were conducted using the National Center for Supercomputing Applications (NCSA) in Illinois, and the University of Arizona offers High Performance Computing (HPC). The study used 418 hours of computation on type gpuA100*4 (TDP of W) hardware. The total estimated emissions are 45.14 kgCO_2eq. According to the United States Environmental Protection Agency United States Environmental Protection Agency (2015), this amount of emissions is equivalent to driving 115 miles in an average gasoline-powered passenger vehicle.

D Annotation Process

As indicated in Osorio et al. (2024), the annotation process involved eight steps:

- First, 12 human coders with domain-specific knowledge in political science and international relations received extensive training on the codebook. These annotators possessed bilingual skills in either English and Spanish, or English and Arabic.
- Second, the coders worked on various sets of randomly sampled 300 aligned sentences. For each set, we had three or four coders. Each human coder processed each individual sentence.
- Third, coders performed a first round of sentence classification blindly. Preventing coders from seeing the annotations conducted by other coders prevents artificial inter-coder correlation. Coders classified each sentence into any of the QuadClass categories or marked them as non-relevant.
- Fourth, after finishing the first round of blind annotations, coders compared their annotations in a non-blind revision round. This helps to rectify discrepancies between coders and strengthen their mastery of the codebook.
- Fifth, sentences with unanimous agreement are considered GSR annotations.
- Sixth, for those sentences in which there was no initial unanimous agreement, coders resolved disagreements in a third round of reviews to enhance inter-coder reliability.
- Seventh, for unresolved sentences, a final coder made the ultimate classification decision.
- Finally, sentences with unresolved classifications or multiple QuadClass labels were excluded from the final dataset.

E Annotation Result

Figures 8 and 9 present the distribution of annotations for the binary classification (relevant or not) and the QuadClass classification, indicating whether a sentence can be categorized as Material Conflict (Mat Conf), Material Cooperation (Mat Coop), Verbal Conflict (Verb Conf), Verbal Cooperation (Verb Coop), or not relevant. For the Binary QuadClass task, the study uses each of the QuadClass as a binary classification.

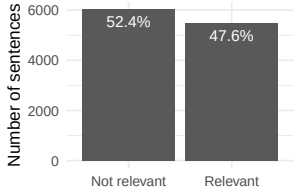


Figure 8: Binary Annotations

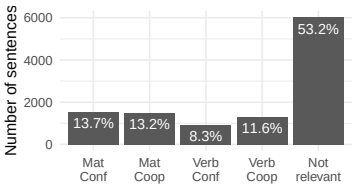


Figure 9: QuadClass Annotations

F MT Tools Training

Each of the MT tools used in this analysis was trained on different corpora of multilingual text, some of which may have included or are known to include multilingual UN documents. While GT and, consequently, Deep, as GT variant, were originally trained on UN documents, they do not specifically include the UNPC as training data (Schäferhoff, 2024). OPUS, in contrast, specifically includes the UNPC as part of its multilingual training corpora, which may result in OPUS showing exceptionally high accuracy in MT the NST text into the target language (OPUS, 2016). DeepL does not specify which training data was used to train the model but emphasizes that DeepL uses a web crawler to find and validate translations on the internet (DeepL, 2024). Consequently, it is possible that DeepL also included UN multilingual documents as training data. While all MT tools can, therefore, be assumed to have been trained on some variant of multilingual UN data, this can be expected to affect the MT output insofar as it is likely to show lower translationese effects than could be expected from an MT model that has not 'seen' the data before when compared to the original UNPC corpus. This limitation notwithstanding, we expect MT tools to differ in terms of their quality and expect their training on UN data not to favor one tool over the others.

G MT Quality Evaluation Metric Configurations

This appendix provides additional technical details related to the configuration used for the MT quality evaluation metrics implemented in the study.

- **SacreBLEU**: Implemented with default settings (tokenizer='13a', force=False, lowercase=False) to ensure reproducibility across multi-lingual settings.
- **METEOR**: Implemented via NLTK library with default settings.
- **BERTScore**: Calculated using bert-base-multilingual-cased model.
- **COMET**: Computed using wmt20-comet-da model with default configurations.

H Nouns, Verbs, and Lemmas

Figure 10 presents the noun count comparison, Figure 11 the verb count comparison, and Figure 12 the lemma count comparison across the native language corpora and MT corpora across languages.

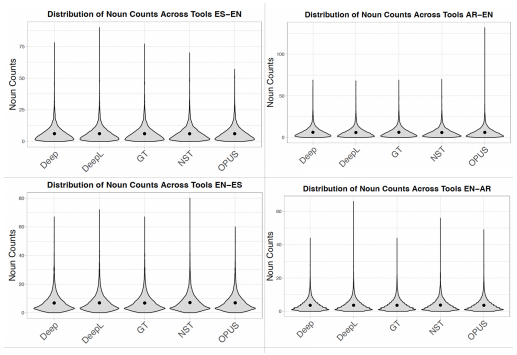


Figure 10: Noun Count Difference

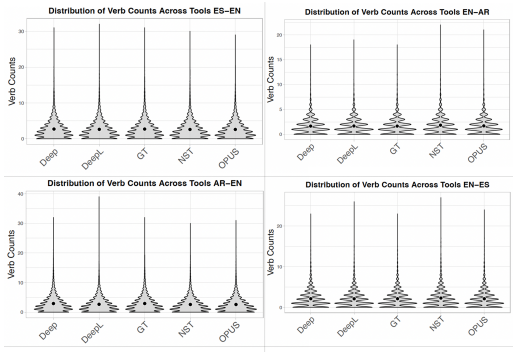


Figure 11: Verb Count Difference

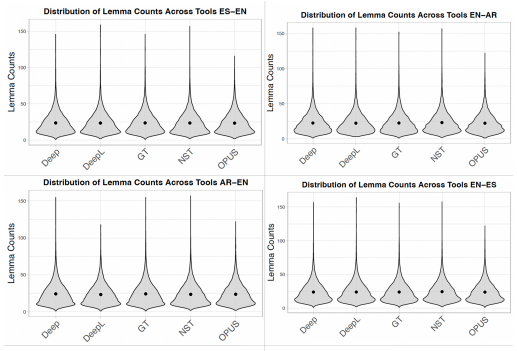


Figure 12: Lemma Count Difference

I Dependency Distance

Figure 13 presents the distribution of the dependency distance mean (DDM), and Figure 14 shows the DDM difference (DDM^d) between the MT texts and their corresponding native language.

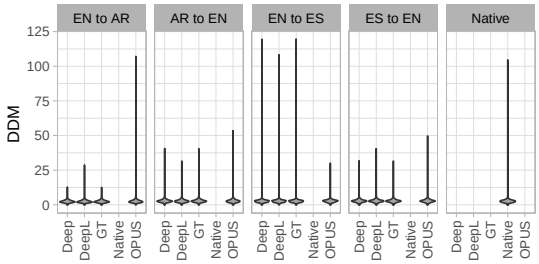


Figure 13: Dependency Distance Mean

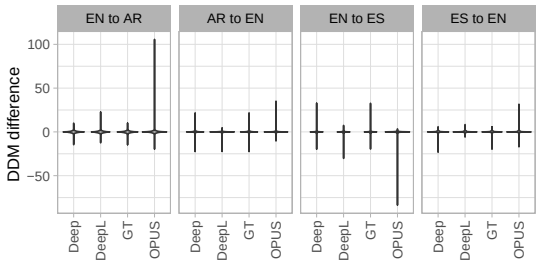


Figure 14: Dependency Distance Mean Difference

J Regression Results

Figure 15 presents the results of the regression analysis indicated in equation 1, where the dependent variable in the probability of ConfliBERT correctly categorizing each sentence in the binary classification task. Coefficients present the point estimate with confidence intervals at 95% of statistical significance. Estimates to the right of the 0 threshold indicate that such sentence characteristic increases the model performance. In contrast, estimates to the left of the threshold indicate a reduction in model performance.

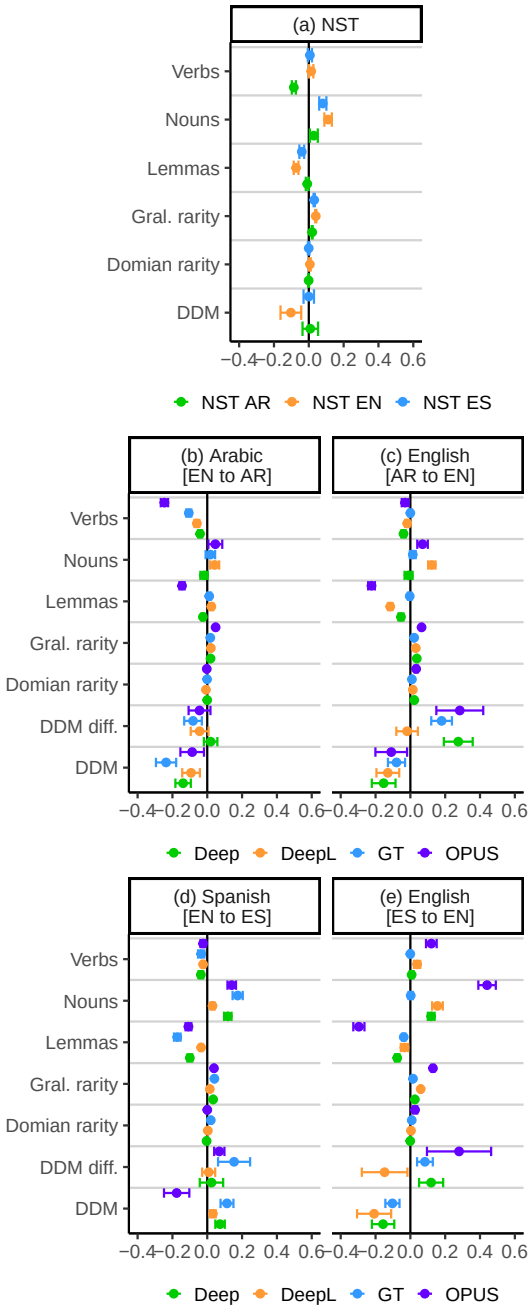
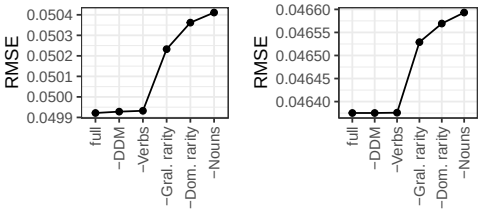


Figure 15: Determinants of Model Performance

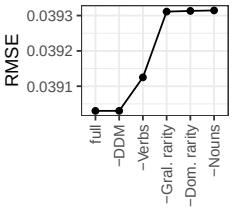
K Individual RMSE plots

The Figures in this Appendix present the original Root Mean Standard Errors (RMSE) generated by each regression. In these plots, the higher RMSE value indicates broader disturbances and, consequently, a lower model fit for ConfliBERT correctly predicting the binary classification task. Figure 16 reports the RMSE from the regressions using the native languages. Figure 17 reports the RMSE from the regressions using the Arabic to English MT output. Figure 18 reports the RMSE from the regressions using the Spanish to English MT text. Figure 19 reports the RMSE from the regressions using the English to Arabic MT documents. Figure 20 reports the RMSE from the regressions using the English to Spanish MT sentences.



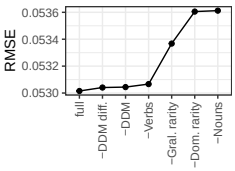
(a) EN NST

(b) ES NST

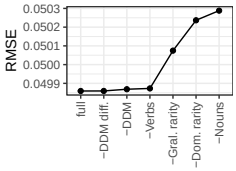


(c) AR NST

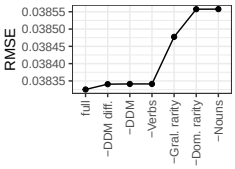
Figure 16: RMSE from NST Text



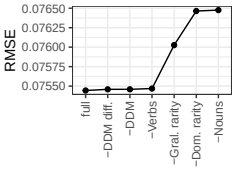
(a) AR to EN Deep



(b) AR to EN DeepL

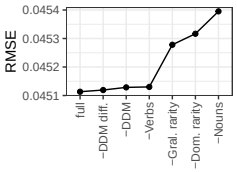


(c) AR to EN GT

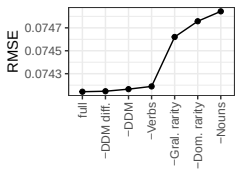


(d) AR to EN OPUS

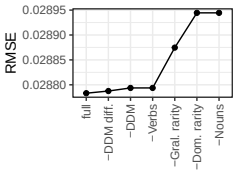
Figure 17: RMSE from Arabic (AR) to English (EN)



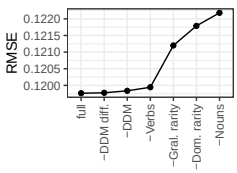
(a) ES to EN Deep



(b) ES to EN DeepL

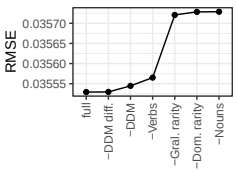


(c) ES to EN GT

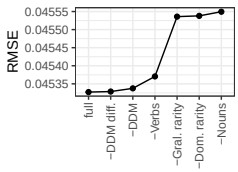


(d) ES to EN OPUS

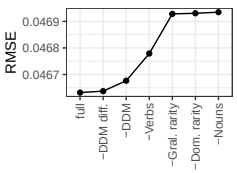
Figure 18: RMSE from Spanish (ES) to English (EN)



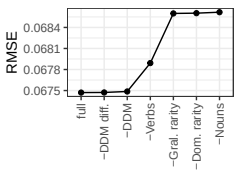
(a) EN to AR Deep



(b) EN to AR DeepL

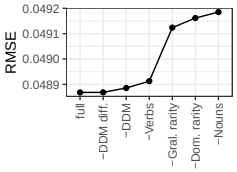


(c) EN to AR GT

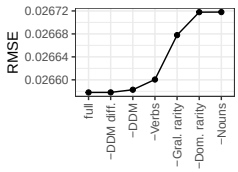


(d) EN to AR OPUS

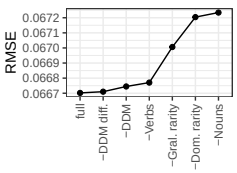
Figure 19: RMSE from English (EN) to Arabic (AR)



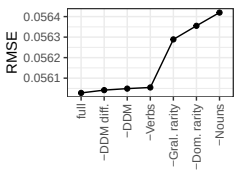
(a) EN to ES Deep



(b) EN to ES DeepL



(c) EN to ES GT



(d) EN to ES OPUS

Figure 20: RMSE from English (EN) to Spanish (ES)

L Acknowledement

This research was supported in part by NSF awards DMS-1737978, DGE-2039542, OAC-1828467, OAC-2311142, and DGE-1906630, ONR awards N00014-17-1-2995 and N00014-20-1-2738, Army Research Office Contract No. W911NF2110032. This work used Delta at NCSA / University of Illinois through allocation CIS220162 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF grants 2138259, 2138286, 2138307, 2137603, and 2138296. Part of the materials presented in this study were processed using the High Performance Computing (HPC) resources supported by the University of Arizona TRIF, UITS, and Research, Innovation, and Impact (RII).

Improving Japanese-English Patent Claim Translation with Clause Segmentation Models based on Word Alignment

Masato Nishimura[1] **Kosei Buma**[1] **Takehito Utsuro**[1] **Masaaki Nagata**[2]
[1]University of Tsukuba [2]NTT Communication Science Laboratories
{s2320779, s2520812}_@_u.tsukuba.ac.jp utsuro_@_iit.tsukuba.ac.jp
masaaki.nagata_@_ntt.com

Abstract

In patent documents, patent claims represent a particularly important section as they define the scope of the claims. However, due to the length and unique formatting of these sentences, neural machine translation (NMT) systems are prone to translation errors, such as omissions and repetitions. To address these challenges, this study proposes a translation method that first segments the source sentences into multiple shorter clauses using a clause segmentation model tailored to facilitate translation. These segmented clauses are then translated using a clause translation model specialized for clause-level translation. Finally, the translated clauses are rearranged and edited into the final translation using a reordering and editing model. In addition, this study proposes a method for constructing clause-level parallel corpora required for training the clause segmentation and clause translation models. This method leverages word alignment tools to create clause-level data from sentence-level parallel corpora. Experimental results demonstrate that the proposed method achieves statistically significant improvements in BLEU scores compared to conventional NMT models. Furthermore, for sentences where conventional NMT models exhibit omissions and repetitions, the proposed method effectively suppresses these errors, enabling more accurate translations.

1 Introduction

The claims in patent documents are critically important for defining the scope of patent rights. However, due to the length and unique descriptive style of these sentences, neural machine translation (NMT) models often encounter issues such as omissions and repetitions in translation. Figure 1 shows the distribution of subword token lengths for Japanese patent claims included in the Japanese-English patent parallel corpus JaParaPat (Nagata et al., 2024a) and for Japanese sentences commonly used in the ASPEC (Nakazawa et al., 2016) Japanese-English parallel corpus. Comparing the two reveals that the patent parallel corpus used in this study has a higher proportion of long sentences compared to scientific paper's abstract. The divide-and-conquer translation approach is known to be an effective method for addressing challenges in long sentences translation. Sudoh et al. (2010) proposed a method in statistical machine translation (SMT) that segments input sentences into clause units based on syntactic parsing, translates each clause separately, and then reorders them according to their hierarchical structure. This approach was shown to improve translation accuracy. Applying this divide-and-conquer approach to neural machine translation (NMT), Kano (2022) proposed a "divide-and-conquer neural machine translation" method for English-Japanese translation, which divides input sentences into clauses based on syntactic parsing and reassembles them after translation. While this method demonstrated the potential to improve translation accuracy, challenges remained in selecting appropriate clause units and ensuring accurate reassembly after clause translation.

In response, Ishikawa (2024) addressed two challenges highlighted in the document (Kano, 2022): the selection of clause segmentation units and the translation accuracy of clauses after segmentation. They sought to improve translation accuracy by adopting clause segmentation based on conjunctions and utilizing mBART (Liu et al., 2020), a pre-trained model, for both the clause translation model and the reordering/editing model. Additionally, they attempted to enhance clause translation accuracy by fine-tuning the clause translation model with pseudo-parallel data at the clause level. Experiments showed a significant reduction in excessively long translations, as well as suppression of

Proceedings of Machine Translation Summit XX Volume 1, pages 333–343, June 23-27, 2025

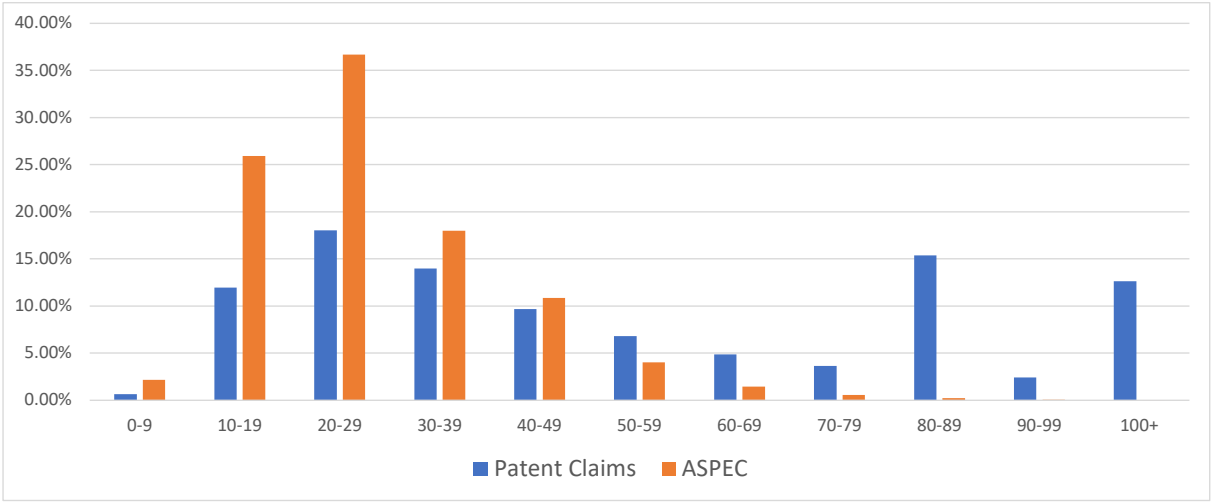


Figure 1: Comparison of Sentence Lengths between ASPEC and Patent Claim Test Data

hallucinations and repetitions. These findings suggest that the divide-and-conquer translation method has the potential to mitigate translation errors commonly caused by conventional NMT models in long sentences translation.

However, the study (Ishikawa, 2024) has some limitations. One issue is that clause segmentation based on conjunctions sometimes fails to divide long sentences into appropriately short clauses. Another issue is that the clause translation model was trained using pseudo-parallel data rather than real parallel data collected from actual sources.

Based on the above, this study proposes a novel approach to the divide-and-conquer translation method, specifically targeting Japanese-English translation of patent claims, which differs from previous studies (Kano, 2022; Ishikawa, 2024). In this method, we introduce a clause segmentation model that divides the source patent claim sentences into clauses optimized for translation by the model. In particular, this study ensures that the clause units, determined based on word alignments in parallel texts, are consistent between the two languages. By doing so, the proposed method suppresses errors such as omissions and repetitions in the final translations, enabling the generation of more accurate translations. Specifically, we propose a method to generate high-quality clause-level parallel data from the original parallel corpus using a word alignment tool. Furthermore, we propose a method to construct the following three models using the generated clause-level parallel corpus:

1. A clause segmentation model that divides the source Japanese sentences into clause units.

2. A clause translation model specialized for clause-level translation.

3. A reordering and editing model that rearranges and edits the translated clauses to generate the final translated text.

In the experiments, the proposed translation method, which integrates these models, was evaluated using the Japanese-English patent parallel corpus JaParaPat (Nagata et al., 2024a). The results demonstrated that the proposed method achieved statistically significant improvements in BLEU scores compared to conventional NMT models for Japanese-English translation of patent claims. Furthermore, compared to the translation results of conventional models, it was confirmed that the proposed method effectively suppresses omissions, resulting in more accurate translations.

2 Related Work

2.1 Long Sentences Translation

Various approaches have been explored to address the challenges of long sentences translation. Sudoh et al. (2010) adopted a divide-and-conquer translation method in statistical machine translation for translating long sentences. They divided input sentences into clause units based on syntactic parsing, translated them, and reordered the results using the hierarchical structure of the clauses, thereby improving translation accuracy.

In NMT, Pouget-Abadie et al. (2014) proposed an automatic segmentation method, which splits long sentences into clauses, translates each clause

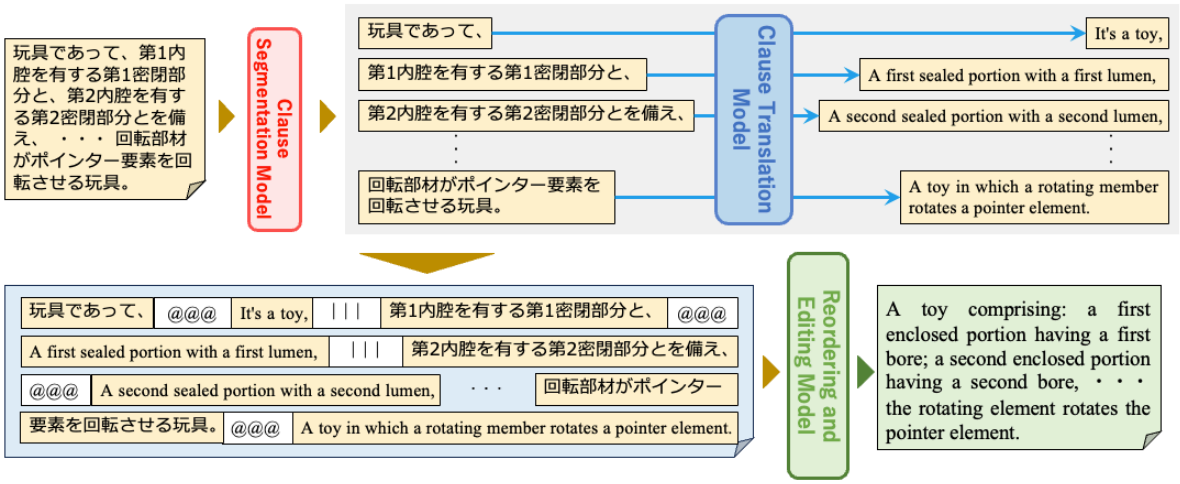


Figure 2: The Prediction Framework of Proposed Model

individually, and then reassembles them sequentially. This method utilizes an RNN to predict the optimal segmentation points for dividing long sentences into parts that are easier for the model to translate. However, this approach was designed for English-French translation. When applied to language pairs with significantly different word order, such as Japanese-English translation, it often resulted in unnatural word order during reassembly.

To address this issue, Kano (2022) developed a neural network model that divides long sentences into smaller segments for translation and rearranges the translated clauses into the appropriate order in English-Japanese translation. Furthermore, Ishikawa (2024) proposed a novel segmentation method for English clauses based on coordinating conjunctions, along with a training method for a model that references the context of the sentence during the translation of the clause. This approach achieved improvements in translation accuracy for long English-Japanese sentences.

2.2 Translation of Patent Claims

Fuji et al. (2015) proposed a method for translating English, Chinese, and Japanese patent claims using statistical machine translation (SMT). Their approach involved manually constructing synchronous context-free grammar rules for sentence structure transformation. These rules were then used to convert the sentence structure of the source language into that of the target language, addressing the unique descriptive style commonly found in patent claims. However, this method has a limitation: the need for manual rule creation makes it difficult to flexibly adapt to new descriptive styles.

3 Method

Figure 2 illustrates the overall structure of the proposed method. The source Japanese patent claim is first divided into multiple clauses using the clause segmentation model. Each clause is then translated by the clause translation model, and finally, the translated clauses are integrated by the reordering and editing model to generate the English patent claim. This method aims to suppress omissions and repetitions that are commonly encountered in conventional NMT models.

It is worth noting that the term *clause* used in this study does not refer to syntactic clauses in the traditional linguistic sense. As shown in Figure 2, in our approach, Japanese sentences are first segmented at punctuation marks, and adjacent segments are then grouped based on word alignments to ensure semantic correspondence with the English side. Thus, we define clauses as semantically coherent segments that preserve consistency between source and target languages. This operational definition aims to support alignment quality rather than adhere strictly to syntactic boundaries.

3.1 Clause-Level Parallel Corpus

In this study, we propose a method for automatically generating a clause-level parallel corpus, inspired by the approach of Zhang and Matsumoto (2019), which generates parallel sub-sentences from long parallel sentence data. This method obtains word alignment information from sentence-level parallel corpora using the word alignment

tool WSPAlign (Wu et al., 2023). Based on the word alignment information, corresponding clauses within sentences are extracted to generate clause-level parallel data.

The clause-level parallel corpus is created using the following procedure. Following the report by Zhang and Matsumoto (2019), we set the word inclusion ratio threshold to 0.5.

1. Use WSPAlign to obtain word alignments for the parallel sentences in the patent parallel data.

2. Split the Japanese and English sentences into multiple clauses at the positions of delimiters such as "、", "，", "。", "．", "；", and "：".

3. Calculate the word inclusion ratio for each pair of parallel clauses based on the word alignment information. If the ratio exceeds 0.5, the clauses are determined to have a alignment. The word inclusion ratio is defined as the proportion of words in a Japanese clause $s\text{-}seg_i$ that are aligned, based on word alignment, to words in the corresponding English clause $t\text{-}seg_j$. In cases where none of the Japanese clauses have a word inclusion ratio larger than 0.5 with any English clause, no clause pairs are created from that sentence pair.

4. For cases where clause alignments are one-to-many or many-to-many, merge the multiple clauses into a single clause on one side to ensure a one-to-one alignment.

By applying the above procedure to parallel sentences extracted from a patent parallel corpus, we generate the clause-level parallel corpus.

architecture	transformer_wmt_en_de_big
enc-dec layers	6
optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$)
learning rate schedule	inverse square root decay
warmup steps	4,000
max learning rate	0.001
dropout	0.3
gradient clip	0.1
batch size	1M tokens
max number of updates	60K steps
validate interval updates	1K steps
patience	5

Table 1: List of hyperparameters for the Transformer

3.2 Clause Segmentation Model

In this study, we developed a clause segmentation model based on ERSATZ, a sentence segmentation model proposed by Wicks and Post (2021). The model was trained to perform segmentation at the clause level. ERSATZ formulates sentence segmentation as a binary classification task, predicting whether periods (e.g., "。" or "．") indicates the "middle of a sentence" or the "end of a sentence". To extend this functionality for clause segmentation, we modified the model to use commas (e.g., "、" or "，") as candidate punctuation marks for clause boundaries.[1] The training data for the model utilized the clause-level parallel corpus proposed in Section 3.1.

The training data was prepared by extracting Japanese clauses from the clause-level parallel corpus and labeling punctuation marks at clause boundaries (e.g., commas) with end-of-clause labels. This enabled the creation of a model capable of segmenting Japanese patent claims into clauses based on word alignment information.

3.3 Clause Translation Model

In this study, to create a clause translation model specialized for clause-level translation, we fine-tuned a pre-trained Japanese-English translation model, initially built using a patent parallel corpus as was also the case in prior studies, with the clause-level parallel corpus generated by the method described in Section 3.1. The experimental settings for the clause translation model, summarized in Table 1, follow those used in JaParaPat (Nagata et al., 2024b). The clause translation model aims to suppress the tendency to infer or supplement contextual information that may be lost due to segmentation, thereby enabling more accurate translations of the segmented clauses.

3.4 Reordering and Editing Model

The purpose of using a reordering and editing model is to reconstruct multiple translated clauses produced by the clause translation model into a single English sentence as the target language sentence. Since the word order in Japanese and English differs significantly, merely dividing a Japanese sentence into clauses and connecting them would not adequately handle the word order

[1] In practice, to perform clause segmentation at positions within parentheses that represent supplementary explanations, the model utilizes both commas (e.g., "、" or "，") and sentence-ending punctuation marks (e.g., "。" or "．").

Model	Data Used	Number of Data
Baseline Model	JaParaPat2016-2020	61,364,685 sentence pairs
Clause Segmentation Model	Clause-Level Parallel Corpus(claims)	200,462 sentence
Clause Translation Model	JaParaPat2016-2019 Clause-Level Parallel Corpus	49,474,547 sentence pairs 5,480,682 clause pairs
Reordering and Editing Model	JaParaPat2016-2020(Bidirectional) JaParaPat2016-2020(claims)	109,028,682 sentence pairs 2,613,107 sentence pairs

Table 2: Overview of Data Used for the Baseline Model and Proposed Method

Evaluation Target	Overall		Long Sentences	
	BLEU ↑	MetricX-24 ↓	BLEU ↑	MetricX-24 ↓
Baseline Model	55.5	2.90	50.1	4.77
Ishikawa	56.3	2.89	51.1	4.76
Proposed Method	**56.6****	**2.84**	**51.6****	**4.69**

Table 3: BLEU Scores and MetricX-24 Scores for Each Evaluation Target. ** indicates a significant difference ($p<0.01$) in BLEU Scores between the Baseline Translation Model and the Proposed Method.

transformation between these languages. Therefore, the reordering and editing model is expected to rearrange the translated clauses into the appropriate word order during the process of connecting them. An example of reordering and editing is shown at the bottom of Figure 2.

The training data for the reordering and editing model is prepared by segmenting sentences in the corpus using the clause segmentation model. The segmented Japanese clauses, along with their translated English clauses, are concatenated to form the input data, while the original English sentences from the corpus are used as the target data. Special tokens, "@@@" and "|||", are added to the model's vocabulary. The token "@@@" is used to connect a Japanese clause with its corresponding translated English clause, while "|||" is used to link pairs of these clause segments. The reason for structuring the input data this way is to preserve information about the relationships between the translated English clauses by including the original Japanese sentence. If only the translated English clauses were used as input, information about the relationships between the clauses would be lost. Adding the Japanese text provides additional context.

Since the input to the reordering and editing model contains words in both Japanese and English, it requires an understanding of both languages. Therefore, the reordering and editing model is created by fine-tuning a Japanese-English bidirectional translation model using the training data prepared as described above.

4 Experiments

4.1 Experimental Setup

In this study, experiments on Japanese-English translation were conducted using the JaParaPat Japanese-English patent parallel corpus (Nagata et al., 2024a). The data used for the experiments consisted of full-text patent parallel data from 2016 to 2020 as the training data, and patent claim parallel data from the first half of 2021 as the test data.

The machine translation software used fairseq (Ott et al., 2019), and the Transformer big (Vaswani et al., 2017) architecture was employed for the baseline model, clause translation model, and reordering and editing model. Sentence tokenization was performed using sentencepiece (Kudo and Richardson, 2018). The model was trained on 10M randomly sampled sentence pairs from the patent parallel data. The vocabulary size was set to 32K for both Japanese and English. Additionally, the clause segmentation model was trained using ERSATZ[2].

Table 2 provides an overview of the data used to train the baseline model and the three proposed models. The clause-level parallel corpus was created by obtaining word alignment information us-

[2] https://github.com/rewicks/ersatz

(a) the Entire Test Set

		Baseline Model		
		2 or more	2–0.5	0.5 or less
Proposed Method	2 or more	1,055	376(*)	4
	2–0.5	273(**)	234,489	900(##)
	0.5 or less	2	515(#)	1,217

(b) the Subset of Inputs with less than 100 Tokens

		Baseline Model		
		2 or more	2–0.5	0.5 or less
Proposed Method	2 or more	651	279(*)	3
	2–0.5	194(**)	202,298	528(##)
	0.5 or less	2	236(#)	695

(c) the Subset of Inputs with 100 to 150 Tokens

		Baseline Model		
		2 or more	2–0.5	0.5 or less
Proposed Method	2 or more	137	41(*)	0
	2–0.5	32(**)	16,155	97(##)
	0.5 or less	0	62(#)	95

(d) the Subset of Inputs with more than 150 Tokens

		Baseline Model		
		2 or more	2–0.5	0.5 or less
Proposed Method	2 or more	265	53(*)	1
	2–0.5	47(**)	15,823	274(##)
	0.5 or less	0	217(#)	424

Table 4: Omission and Repetition Analysis (Proposed Method vs. Baseline Model) on the Entire Test Set

ing WSPAlign[3] for half of the 2020 data (5,976,295 sentence pairs) and following the method described in Section 3.1. This process resulted in a clause-level parallel corpus containing 5,480,682 clause pairs. For training the clause segmentation model, Japanese clause data was created by segmenting Japanese patent claims in the clause-level parallel corpus. The clause translation model was pre-trained on the full-text patent parallel data from 2016 to 2019 and fine-tuned using the entire clause-level parallel corpus. The reordering and editing model was pre-trained on bidirectional full-text patent parallel data from 2016 to 2020 and fine-tuned using training data created by applying the methods described in Section 3.4 to Japanese patent claims from 2016 to 2020, segmented and translated using the clause segmentation and translation models.

To compare with conventional divide-and-conquer neural machine translation methods, we reproduced Ishikawa (2024)'s approach, which involves clause segmentation based on conjunctions and fine-tuning a clause translation model with pseudo-parallel data at the clause level, adapting it for Japanese-to-English translation of patent claims. The training data used for this reproduction was within the same range as the data used to train the three models in the proposed method. This comparison allows us to evaluate the effectiveness of using the clause segmentation model adopted in the proposed method and the clause-level parallel corpus created using word alignment information.

For evaluation, BLEU (Papineni et al., 2002) was

[3] https://github.com/qiyuw/WSPAlign

used as the primary metric, calculated with sacreBLEU[4] (Post, 2018). Since accurate translation of technical terms is critical in patent translation, BLEU was selected as the main evaluation criterion in this study.

To evaluate whether the proposed method can suppress translation errors such as omissions and repetitions, we conducted an assessment using MetricX-24[5] (Juraska et al., 2024). MetricX-24 is a machine translation evaluation metric developed by Google based on a regression model that predicts MQM (Multidimensional Quality Metrics) scores (Lommel et al., 2014). Traditional machine translation evaluation metrics, such as COMET (Rei et al., 2022a) and CometKiwi (Rei et al., 2022b), are trained on Direct Assessment (DA) scores and are highly effective in measuring semantic adequacy. However, MQM scores allow for weighting different types of translation errors, making them more suitable for evaluating issues such as omissions and repetitions. Furthermore, MetricX-24 has demonstrated high robustness against translation errors, including omissions and repetitions, by leveraging mixed training on synthetic error data and DA/MQM data. In this study, we used the MetricX-24-Hybrid-XL[6] model for evaluation.

4.2 Results

4.2.1 Accuracy Evaluation

In this study, the performance of the proposed method was evaluated using a test set consisting of patent claims (238,902 sentences) extracted from 2021 patent data. Table 3 showed that the proposed method achieved a BLEU score of 56.6, which statistically significantly outperformed the baseline model's score of 55.5 ($p < 0.01$). This confirmed that the proposed method improves overall translation accuracy.

Additionally, the performance was evaluated on a subset of the test set containing only long sentences with more than 100 subword tokens in the source Japanese text. For this subset, the proposed method recorded a BLEU score of 51.6, statistically significantly exceeding the baseline model's score of 50.1. While the overall test set showed an improvement of 1.1 points, the improvement for long sentences was 1.5 points, indicating that the proposed method achieved greater improvement for longer sentences.

The results using MetricX-24 showed that the proposed method achieved a score of 2.84, compared to 2.90 for the baseline model. In MetricX-24, lower scores indicate fewer translation errors, such as omissions and repetitions. This suggests that the proposed method effectively suppresses translation errors in patent claim translations, including omissions and repetitions.

Next, a comparison was made between the proposed method and conventional divide-and-conquer neural machine translation methods. For the conventional method, the BLEU score was 56.3 points, and the MetricX-24 score was 2.89 points. In contrast, the proposed method achieved a BLEU score of 56.6 and a MetricX-24 score of 2.84, outperforming the conventional method in both metrics. Both the baseline and our proposed method use parallel data extracted from the same portion of JParaPat.Our method differs from previous divide-and-conquer approaches in a key aspect: whereas prior methods typically rely solely on the syntactic structure of the source language—often segmenting at coordinating conjunctions—our proposed approach leverages word alignments to identify clause boundaries based on source–target correspondence. This alignment-based segmentation results in divisions that are more suitable for translation. These results confirm that, compared to the conventional divide-and-conquer neural machine translation method, the clause segmentation model and the clause-level parallel corpus leveraging word alignment information employed in the proposed method contribute to improved accuracy in divide-and-conquer neural machine translation.

4.2.2 Analysis of Omissions and Repetitions

To further analyze whether the proposed method can produce more accurate translations with fewer errors such as omissions and repetitions, the sentence length ratios between the translated text and the reference text were calculated for both the baseline model and the proposed method. These ratios were categorized into three groups: "2 or more," "2–0.5," and "0.5 or less," and their trends were observed. The classification results for the entire test set (238,902 sentences) are shown in Table 4 (a). Additionally, the test set was grouped by the token length of the input sentences into "less than 100," "100–150," and "more than 150," with the classification results for each group shown in Ta-

[4] https://github.com/mjpost/sacrebleu
[5] https://github.com/google-research/metricx
[6] https://huggingface.co/google/metricx-24-hybrid-xl-v2p6

bles 4 (b), 4 (c), and 4 (d), respectively. Within these tables, special attention was given to the four categories where one model produced translations with omissions or repetitions while the other model performed well: "middle column, upper row(*)," "left column, middle row(**)," "middle column, lower row(#)," and "right column, middle row(##)."

The analysis showed that in all four tables, the "left column, middle row(**)," which represents cases where the proposed method successfully avoided repetitions that the baseline model did not, occurred less often than the "middle column, upper row(*)." This suggests that the baseline model had fewer cases of repetition overall. On the other hand, for omissions, the "middle column, lower row(#)," where the proposed method avoided omissions, occurred less frequently than the "right column, middle row(##)." This indicates that the proposed method was better at reducing omissions compared to the baseline model.

Examples of improved translations addressing omissions by the proposed method are shown in Table 6. The baseline model in Table 6 (a), parts of the input sentence, such as "preferably by a length of the heat exchanger" and "finned tube shape, coiled shape, and/or fin shape", were not translated despite being present in the original Japanese sentence. Additionally, in patent claims, reference numerals in drawings are typically enclosed in parentheses, as seen in "The cryogenic refrigeration system (1)". However, in the baseline model's translation, the number inside the parentheses was omitted. In contrast, the proposed method not only translates the entire input Japanese sentence without missing any information but also correctly retains the numerical references within parentheses. As a result, it produces a more appropriate translation for patent claims. Similarly, in Table 6 (b), the baseline model fails to translate some words in input sentence such as "such as methanol, ethanol," whereas the proposed method correctly translates all examples. These results indicate that, compared to the baseline model, the proposed method preserves all necessary information in patent claim translations and produces more accurate outputs.

Examples of improved translations addressing repetitions by the proposed method are shown in Table 7. In the baseline model, the term "cantilever shaped" was excessively repeated, whereas no such repetition occurred with the proposed method.

4.2.3 Impact of Pre-training the Reordering and Editing Model

In this study, bidirectional Japanese-English parallel data was used for pre-training the reordering and editing model. Experiments were conducted to evaluate the effect of this pre-training on the accuracy of the final reordering and editing model. The parallel data used for pre-training consisted of JaParaPat data from 2016 to 2020, and two models were created: one trained with Japanese-English parallel data and the other with bidirectional parallel data. These models were fine-tuned using the same reordering and editing model training data, and their performance was compared.

The BLEU evaluation results, obtained using test data comprising patent claims extracted from 2021 patent data, are shown in Table 5. The results show that the reordering and editing model pre-trained with bidirectional Japanese-English parallel data achieved a BLEU score of 56.6, statistically significantly outperforming the model pre-trained only in the Japanese-to-English direction, which scored 55.0 ($p < 0.01$). The results suggest that understanding both Japanese and English is critical for the reordering and editing model. Furthermore, using bidirectional Japanese-English parallel data for pre-training improves the accuracy of reordering and editing.

Table 5: BLEU scores of the Reordering and Editing Model with different pre-training data: comparison between Unidirectional (Japanese-English) and Bidirectional (Japanese-English) parallel data. ** indicates a significant difference (p<0.01) in BLEU scores.

Data used for Pre-Training	BLEU
Unidirectional	55.0
Bidirectional	**56.6****

4.2.4 Evaluation of the Clause Segmentation Model

The clause segmentation model developed in this study was evaluated to determine its ability to accurately segment Japanese patent claim sentences. For the evaluation, 2,000 sentences were sampled from 238,902 patent claim sentences extracted from 2021 patent data. First, word alignment information was obtained for the 2,000 sentences using WSPAlign, and the sentences were segmented into clauses based on the method described in Section 3.1. Ground truth data was then created by assigning end-of-sentence labels to punctuation

(a) Example 1

Omissions

Input Sentence
前記温度因子及び/又は前記NTUが、前記熱交換器(3)の伝熱面積によって、好ましくは前記熱交換器の長さによって提供され、前記熱交換器(3)が、好ましくは、フィン付きチューブ形状、コイル形状、及び/又はフィン形状であり、前記流路(2)の円周を少なくとも部分的に取り囲む、 請求項2に記載の極低温冷凍システム(1) 。

Reference Translation
Cryogenic refrigeration system (1) according to claim 2, wherein the temperature factor and/or the NTU is provided by a heat transfer area of the heat exchanger (3), preferably by a length of the heat exchanger, wherein the heat exchanger (3) is preferably of a finned tube shape, coiled shape, and/or fin shape and at least partially surrounds a circumference of the conduit (2).

Baseline Model (BLEU: 15.65, COMET: 68.94, MetricX-24: 2.82)
The cryogenic refrigeration system according to claim 2, wherein the temperature factor and/or the NTU is provided by a heat transfer area of the heat exchanger.

Proposed Method (BLEU: 78.24, COMET: 84.31, MetricX-24: 2.12)
The cryogenic refrigeration system (1) according to claim 2, wherein the temperature factor and/or the NTU is provided by a heat transfer area of the heat exchanger (3), preferably by a length of the heat exchanger, wherein the heat exchanger (3) is preferably finned tube-shaped, coil-shaped and/or fin-shaped and at least partially surrounds a circumference of the flow channel (2).

(b) Example 2

Omissions

Input Sentence
有機溶媒1が、アルコール溶媒、例えば、メタノール、エタノール、n−プロパノール、イソプロパノール、n−ブタノール、イソブタノール;エステル溶媒、例えば、酢酸メチル、酢酸エチル、酢酸プロピル、酢酸イソプロピル、酢酸ブチル;ケトン溶媒、例えば、アセトンおよびブタノン;またはその混合物である、請求項9に記載の方法。

Reference Translation
The method according to claim 9, wherein the organic solvent 1 is an alcohol solvent, such as methanol, ethanol, n-propanol, isopropanol, n-butanol, isobutanol; an ester solvent, such as methyl acetate, ethyl acetate, propyl acetate, isopropyl acetate, butyl acetate; a ketone solvent, such as acetone and butanone; or a mixture thereof.

Baseline Model (BLEU=21.11, COMET=60.85, MetricX-24=2.74)
The method according to claim 9, wherein the organic solvent 1 is an alcohol solvent; an ester solvent; a ketone solvent; or a mixture thereof.

Proposed Method (BLEU=88.99, COMET=90.98, MetricX-24=1.77)
The method according to claim 9, wherein the organic solvent 1 is an alcohol solvent such as methanol, ethanol, n-propanol, isopropanol, n-butanol, isobutanol; an ester solvent such as methyl acetate, ethyl acetate, propyl acetate, isopropyl acetate, butyl acetate; a ketone solvent such as acetone and butanone; or a mixture thereof.

Table 6: Examples of Omission Improvements by the Proposed Method

marks at the segmentation points, which served as the test set for evaluating the accuracy of the clause segmentation model. Next, the clause segmentation model created in Section 4.1 was applied to the test set's Japanese patent claim sentences to perform clause segmentation. The segmentation points predicted by the model were compared to the ground truth segmentation points, and the F1

Repetitions

Input Sentence

前記流路遮断バルブは、 内部に前記閉鎖部材を収容し、カンチレバー形状からなる少なくとも1つの片持ちばりを備え、円筒形に形成されて、前記連通流路の流入口に挿入されるように設置されるボディー;及び、 一側は前記片持ちばりから突出形成される係止部により支持され、他側は前記閉鎖部材と接触するように設置されるリング部材;を含み、 前記リング部材は設定された温度以上になると、前記閉鎖部材が中心部を通過するように変形されて、前記閉鎖部材を前記連通流路の内部に向けて移動させる、 ことを特徴とする、請求項17に記載のバルブアセンブリ。

Reference Translation

The valve assembly of claim 17, wherein the flow path blocking valve includes:a body for containing the blocking member therein, and providing at least one cantilever portion formed in a cantilever shape, the body formed in a cylindrical shape and disposed to be inserted into the inlet of the communication flow path; anda ring member having one side supported by a locking portion protruding from the cantilever portion and the other side disposed to contact the blocking member,wherein the ring member moves the blocking member towards the inside of the communication flow path by deforming the blocking member to pass through a central part of the ring member when the internal temperature exceeds the preset temperature.

Baseline Model (BLEU=7.63, COMET=34.41, MetricX-24=8.48)

The valve assembly of claim 17, wherein the flow path shut-off valve comprises:a body configured to accommodate the closing member therein, having at least one cantilever shaped

Proposed Method (BLEU=43.72, COMET=81.22, MetricX-24=4.82)

The valve assembly of claim 17, wherein the passage shutoff valve comprises:a body accommodating the closing member therein, having at least one cantilever formed in a cantilever shape, formed in a cylindrical shape, and installed to be inserted into the inlet of the communication passage; anda ring member having one side supported by a locking portion formed to protrude from the cantilever and the other side installed to be in contact with the closing member,wherein the ring member is deformed such that the closing member passes through a central portion and moves the closing member toward the inside of the communication passage when a set temperature or higher is reached.

Table 7: Examples of Repetition Improvements by the Proposed Method

score was calculated. The model achieved an F1 score of 98.16. These results demonstrate that the clause segmentation model developed in this study can accurately reproduce clause segmentation by effectively utilizing word alignment information.

5 Conclusion

In this study, we proposed a translation method to address the translation errors of “omissions” and “repetitions” which are common challenges in Japanese-English translation of patent claims. The method focuses on the fact that patent claims are often long and have unique structures, utilizing a clause segmentation model to divide patent claims into more translatable units.

The experimental results demonstrated that the proposed method achieved statistically significant improvements over the baseline model in BLEU scores. Notably, it showed remarkable improvements even for sentences prone to omissions and repetitions. These results confirm the effectiveness of the proposed method in resolving issues of omissions and repetitions in the translation of patent claims.

An important direction for future work is to extend the evaluation to include state-of-the-art large language models (LLMs), which we plan to pursue in our subsequent research.

References

Masaru Fuji, Atsushi Fujita, Masao Utiyama, Eiichiro Sumita, and Yuji Matsumoto. 2015. Patent claim translation based on sublanguage-specific sentence structure. In *Proceedings of Machine Translation Summit XV: Papers*, Miami, USA.

R. Ishikawa. 2024. Divide-and-conquer neural machine translation with insentence context. *Master's Thesis, Graduate School of Science and Technology, Nara Institute of Science and Technology*.

J. Juraska, D. Deutsch, M. Finkelstein, and M. Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proc. 9th WMT*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Y. Kano. 2022. Improving neural machine translation by syntax-based segmentation. *Master's Thesis, Graduate School of Science and Technology, Nara Institute of Science and Technology*.

T. Kudo and J. Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *In Proc. EMNLP*, pages 66–71.

A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Y. Liu et al. 2020. Multilingual denoising pre-training for neural machine translation. *TACL*, 8:726–742.

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

M. Nagata, M. Morishita, K. Chousa, and N. Yasuda. 2024a. JaParaPat: A large-scale Japanese-English parallel patent application corpus. In *Proc. LREC-COLING*, pages 9452–9462.

M. Nagata, M. Morishita, K. Chousa, and N. Yasuda. 2024b. JaParaPat: A large-scale Japanese-English parallel patent application corpus. In *Proc. LREC-COLING 2024)*, pages 9452–9462.

T. Nakazawa, M. Yaguchi, K. Uchimoto, M. Utiyama, E. Sumita, S. Kurohashi, and H. Isahara. 2016. ASPEC: Asian scientific paper excerpt corpus. *In LREC2016*, pages 2204–2208.

M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*, pages 48–53.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. 40th ACL*, pages 311–318.

M. Post. 2018. A call for clarity in reporting BLEU scores. In *Proc. 3rd WMT*, pages 186–191.

J. Pouget-Abadie, D. Bahdanau, B. van Merrien-Boer, K. Cho, and Y. Bengio. 2014. Overcoming the curse of sentence length for neural machine translation using automatic segmentation. *In Proc. 8th SSST*, pages 78–85.

R. Rei, J. C. de Souza, D. Alves, C. Zerva, A. Farinha, T. Glushkova, A. Lavie, L. Coheur, and A. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proc. 7th WMT*, pages 578–585.

R. Rei et al. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proc. 7th WMT*, pages 634–645.

K. Sudoh, K. Duh, H. Tsukada, T. Hirao, and M. Nagata. 2010. Divide and translate: Improving long distance reordering in statistical machine translation. In *Proc. 5th WMT*, pages 418–427.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Proc. 30th NIPS*, pages 5998–6008.

R. Wicks and M. Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. In *In Proc. 59th ACL*, pages 3995–4007.

Q. Wu, M. Nagata, and Y. Tsuruoka. 2023. WSPAlign: Word alignment pre-training via large-scale weakly supervised span prediction. In *Proc. 61st ACL*, pages 11084–11099.

J. Zhang and T. Matsumoto. 2019. Corpus augmentation for neural machine translation with chinese-japanese parallel corpora. *Applied Sciences*, 9(10).

A Sustainability Statement

A.1 CO2 Emission Related to Experiments

Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.432 $kgCO_2eq/kWh$. A cumulative of 1,000 hours of computation was performed on hardware of type RTX A6000 (TDP of 300W).

Total emissions are estimated to be 129.6 $kgCO_2eq$ of which 0 percents were directly offset.

Estimations were conducted using the Machine-Learning Impact calculator presented in Lacoste et al. (2019).

Progressive Perturbation with KTO for Enhanced Machine Translation of Indian Languages

Yash Bhaskar[1], Ketaki Shetye[1], Vandan Mujadia[1], Dipti Misra Sharma[1], Parameswari Krishnamurthy[1]

[1]International Institute of Information Technology – Hyderabad
Correspondence: yash.bhaskar@research.iiit.ac.in, ketaki.shetye@research.iiit.ac.in, vandan.mu@research.iiit.ac.in, dipti@iiit.ac.in, param.krishna@iiit.ac.in

Abstract

This study addresses the critical challenge of data scarcity in machine translation for Indian languages, particularly given their morphological complexity and limited parallel data. We investigate an effective strategy to maximize the utility of existing data by generating negative samples from positive training instances using a progressive perturbation approach. This is used to align the model with preferential data using Kahneman-Tversky Optimization (KTO). Comparing it against traditional Supervised Fine-Tuning (SFT), we demonstrate how generating negative samples and leveraging KTO enhances data efficiency. By creating rejected samples through progressively perturbed translations from the available dataset, we fine-tune the Llama 3.1 Instruct 8B model using QLoRA across 16 language directions, including English, Hindi, Bangla, Tamil, Telugu, and Santali. Our results show that KTO-based preference alignment with progressive perturbation consistently outperforms SFT, achieving significant gains in translation quality with an average BLEU increase of 1.84 to 2.47 and CHRF increase of 2.85 to 4.01 compared to SFT for selected languages, while using the same positive training samples and under similar computational constraints. This highlights the potential of our negative sample generation strategy within KTO, especially in low-resource scenarios.

1 Introduction

Machine Translation (MT) has made remarkable progress in recent years, yet significant challenges persist, particularly for low-resource languages. This is evident in the diverse family of Indian languages, such as Tamil, with its agglutinative morphology (Sarveswaran et al., 2021) and complex

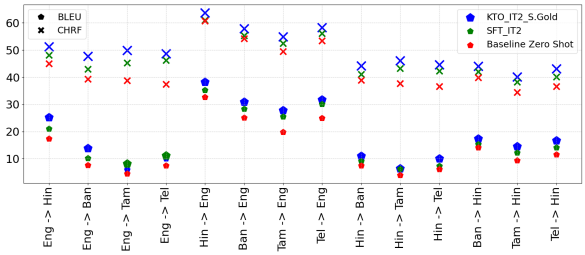


Figure 1: Performance Comparison of KTO (with Progressive Perturbation, using IndicTrans2 (IT2) output as positive samples and Perturbed Gold translations from BPCC Dataset (S.Gold) as negative samples) vs. SFT (using IT2 output as positive samples) and Zero-Shot on Llama 3.1 Instruct 8B.

suffixation, and Santali, which employs an Austroasiatic script (Choksi, 2018) and follows an SOV word order. These languages feature rich morphological systems that complicate tokenization and alignment in MT (Kumar et al., 2009) while also suffering from a scarcity of parallel corpora essential for training robust translation models.

The imbalance in training data between high-resource and low-resource languages has motivated the search for data-efficient techniques that maximize the utility of scarce resources. In this study, we tackle this challenge for Indian language machine translation by leveraging an approach based on preference alignment (Gisserot-Boukhlef et al., 2024). Rather than requiring extra positive training data, our method utilizes negative samples derived from existing high-quality translations. This enables the model to learn more effectively by distinguishing between subtle errors and accurate translations, thereby enhancing overall performance even in resource-scarce settings.

KTO (Ethayarajh et al., 2024) distinguishes itself from other preference-based methods by its flexibility in handling negative samples. Unlike Direct Preference Optimization (DPO) (Mecklenburg et al., 2024), which ideally requires rejected

Proceedings of Machine Translation Summit XX Volume 1, pages 344–352, June 23-27, 2025

completions for each positive example, and Proximal Policy Optimization (PPO) (Schulman et al., 2017), which necessitates the training of a separate and computationally intensive reward model, KTO allows for the utilization of negative samples without demanding a one-to-one pairing with every positive instance. This flexibility is particularly advantageous in low-resource settings, where generating a large number of diverse negative samples is challenging, and fine-tuning them increases computational cost.

In this work, we propose a progressive perturbation strategy to generate negative samples by systematically adding controlled noise to positive translations. These rejected samples, along with the original positives, are then used with the KTO algorithm for preference alignment. This approach enhances translation quality without requiring additional parallel data, making it particularly effective in low-resource scenarios.

We validate our approach on the Llama 3.1 Instruct 8B model across 16 language directions involving English, Hindi, Bangla, Tamil, Telugu, and Santali. Experimental results demonstrate that our KTO-based preference alignment with progressive perturbation consistently outperforms traditional SFT (Ouyang et al., 2022), yielding significant improvements in both BLEU and CHRF scores.

2 Related Work

Low resource machine translation (MT) remains a persistent challenge, motivating a variety of strategies to maximize data efficiency. Early work demonstrated that careful structuring of training data can significantly impact convergence and overall translation quality. For example, (Platanios et al., 2019) introduced a competence-based curriculum that adapts the complexity of training examples to the model's evolving capabilities. In a similar vein, (Zhang et al., 2018) and (Liu et al., 2020) showed that progressively increasing data complexity by ordering training examples from simple to complex can lead to faster convergence and improved performance in MT.

In addition to curriculum learning, data augmentation techniques have been widely explored to overcome the scarcity of parallel corpora in low-resource settings. (Xia et al., 2019) augmented training data using monolingual corpora from related high-resource languages, thereby enriching the available signal without the need for additional bilingual data. Similarly, (Ramesh et al., 2021) proposed a method that leverages bilingual word embeddings and transformer-based representations (e.g., BERT (Devlin et al., 2019)) to introduce new words and increase the presence of rare vocabulary items in the training corpus. While effective, these approaches typically require access to supplementary resources or complex augmentation pipelines.

Data quality also plays a critical role in MT, particularly when dealing with automatically generated or noisy datasets. To address this, (Kowtal et al., 2024) developed a data selection method that uses cross-lingual sentence representations derived from a multilingual SBERT model (Reimers, 2019) to filter out semantically mismatched sentence pairs. This filtering enhances the reliability of the training data but does not directly tackle the challenge of making optimal use of the available examples.

Multilingual transfer learning offers another avenue for improving low-resource MT by exploiting the inherent relatedness between languages. (Goyal et al., 2020) combined techniques such as unified transliteration and shared subword segmentation with pre-training across multiple languages to enhance transfer learning capabilities. Although effective, such approaches generally require a joint training framework that spans multiple language pairs.

In contrast to these paradigms, our work adopts a preference-based optimization strategy that directly maximizes the utility of existing data. Instead of relying solely on positive examples or external augmentation, we generate informative negative samples through a progressive perturbation strategy. By systematically degrading high-quality translations, our approach creates rejected samples that force the model to learn fine-grained distinctions between accurate and flawed outputs.

3 Methodology

We opted to carry out our experiments across six distinct languages divided into three categories as listed below, originating from three to four different language families and varying in resource availability.

1. **English to Indian Languages**: Translations from English to Bangla, Hindi, Santali, Tamil, and Telugu.

2. **Indian Languages to English**: Translations

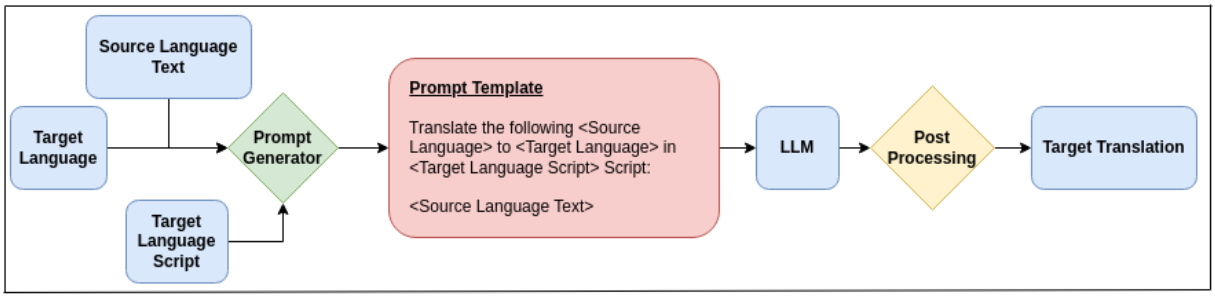


Figure 2: Prompting Mechanism for Translation

from Bangla, Hindi, Santali, Tamil, and Telugu to English.

3. **Indian to Indian Languages**: Translations between Hindi and Bangla, Tamil, and Telugu (excluding Santali due to limited parallel data).

To address data scarcity in Indian language translation, we compare zero-shot inference (baseline), SFT, and KTO. All experiments use the llamaFactory toolkit [1] (Zheng et al., 2024).

3.1 Model Selection

For this study, we selected the Llama 3.1 Instruct 8B model [2] (Dubey et al., 2024) as the foundation for fine-tuning. This choice was made after conducting initial zero-shot experiments to assess the baseline translation performance of several models relevant to our tasks. Specifically, we evaluated the Llama 3.1 Instruct 8B, Llama 3.2 Instruct 3B, and Llama 3.2 Instruct 11B models in a zero-shot setting across the language directions.

Table 1 summarizes the average BLEU and CHRF scores for each model across the language directions, evaluated on the Flores-200 devtest set.

Table 1: Average Zero-Shot BLEU and Chrf Scores for Llama Models

Model	BLEU	Chrf
Llama3.1-8B	12.06	39.00
Llama3.2-3B	5.58	33.22
Llama3.2-11B	11.94	39.47

As evident from Table 1, the Llama 3.2 Instruct 3B model demonstrated significantly lower translation quality compared to both the 8B and 11B parameter versions. Notably, the zero-shot translation performance of the Llama 3.1 Instruct 8B and Llama 3.2 Instruct 11B models was remarkably similar. Given this performance parity, and considering computational resource constraints for extensive fine-tuning experiments, we opted to proceed with the Llama 3.1 Instruct 8B model.

3.2 Data and Preprocessing

To ensure diverse and representative training data, we utilized the Wiki and Massive datasets from the Bharat Parallel Corpus Collection (BPCC) (Gala et al., 2023), sampling data as detailed in Table 2. The languages involved are Bangla (Bengali script), Hindi (Devanagari script), Santali (Ol Chiki script), Telugu (Telugu script), English (Latin script), and Tamil (Tamil script).

Language Pairs	Sample Size
Eng ↔Hin	25,000
Eng ↔Ban	25,000
Eng ↔Tam	25,000
Eng ↔Tel	25,000
Eng ↔San	25,000
Hin ↔Ban	10,000
Hin ↔Tam	10,000
Hin ↔Tel	10,000

Table 2: Language Pairs and Sample Sizes. In this, **Eng** refers to **English**, **Hin** refers to **Hindi**, **Ban** refers to **Bangla**, **Tam** refers to **Tamil**, **Tel** refers to **Telugu**, **Sat** refers to **Santali**

Throughout our experiments, a consistent prompt format was maintained for all techniques to ensure comparability. This prompt structure, visualized in Figure 2, includes specifications for the source and target languages, the target script, and the source sentence for translation.

[1] https://github.com/hiyouga/LLaMA-Factory
[2] https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct

Perturbation Level	Example Sentence
10%	**Original:** A person with proliferative retinopathy will always be at risk for complications from new bleeding as well as glaucoma, new blood vessels. **Perturbed:** A bleeding with proliferative retinopathy will always be at risk for eagle from new person as well as glaucoma, new blood vessels.
30%	**Original:** The confluence of the Mudirapuzha, Nallathani, and Kundala rivers takes place in the heart of the city. **Perturbed:** The confluence of nervosa Mudirapuzha, kiskindha Nallathani, and excreted Kundala takes place in the heart of the city.
50%	**Original:** Most of the street children in Bangalore have come in search of business and new beginnings. **Perturbed:** Most of the children Bangalore in street have come in search of business photos car new.

Table 3: Examples of english sentence perturbations at 10%, 30%, and 50% intensity levels.

3.3 Perturbation Strategy

To introduce controlled errors, we apply a set of text modification operations that simulate common translation errors:

- **Word Addition**: Randomly inserts a word from a predefined vocabulary, disrupting fluency and potential meaning.
- **Word Deletion**: Removes a random word, leading to grammatical errors and incomplete sentences.
- **Word Shuffling**: Swaps the position of two random words, disrupting word order and comprehensibility.
- **Word Replacement**: Replaces a random word with another vocabulary word, introducing semantic errors.

The number of modifications depends on the perturbation intensity level. For instance, at 30% perturbation, a 20-word sentence undergoes approximately six modifications. This progressive perturbation (50% $\rightarrow$ 30% $\rightarrow$ 10%) exposes the model to coarse-to-fine errors, aligning with its improving discrimination capability during training. Some of the examples depicting the different levels of intensity-perturbation can be seen in Table 3

We integrate a *progressive perturbation strategy* with KTO to enhance model training. This method systematically introduces controlled noise into gold-standard human translations and IndicTrans2 (IT2) outputs, generating rejected completions for preference alignment. Perturbations are applied at varying intensities (10%, 30%, 50%), beginning with highly degraded (50%) translations to establish clear negative examples, then progressively reducing perturbation levels (30%, 10%) to introduce more nuanced errors. This staged approach refines the model's ability to distinguish subtle translation flaws, improving overall translation quality.

4 Fine-tuning and Optimization

We compare SFT with KTO, both applied to the Llama 3.1 Instruct 8B model.

4.1 Supervised Fine-Tuning

SFT serves as our baseline, evaluating standard supervised learning with limited parallel data. We explore two variations:

- **SFT on Gold-Standard Translations**: Fine-tuning on a subset of the Massive and Wiki datasets from BPCC using human translations as ground truth, setting a benchmark for high-quality supervision.
- **SFT on IT2-Generated Translations**: Fine-tuning with IT2-generated translations (Gala et al., 2023) as targets, assessing whether synthetic data can supplement or replace human translations in low resource settings.

These variations help assess the impact of different supervision sources on translation performance.

4.2 Kahneman-Tversky Optimization

We evaluate KTO using four configurations to analyze how different data sources influence preference alignment:

- **KTO-Gold-S.IT2**: Gold-standard translations as preferred examples, with rejected samples from perturbed IT2 outputs.
- **KTO-Gold-S.Gold**: Both preferred and rejected examples from gold-standard translations, with perturbation applied for rejection samples.

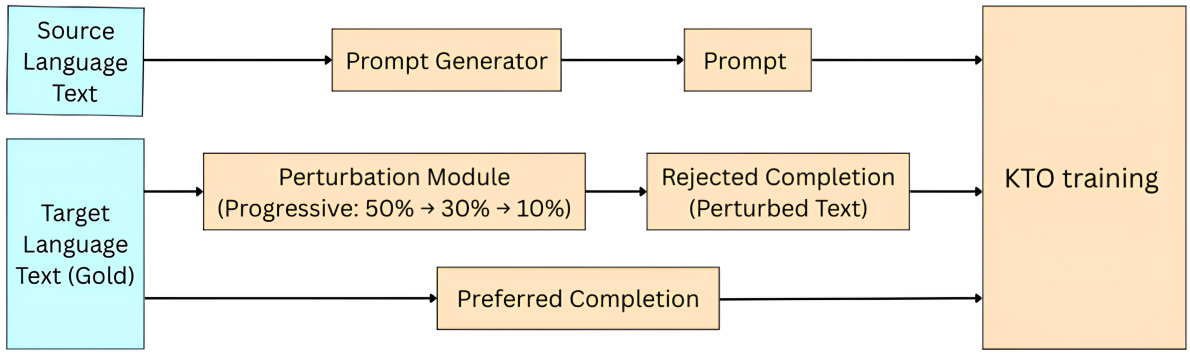


Figure 3: KTO training data workflow using gold text and progressive perturbation.

- **KTO-IT2-S.IT2**: IT2 generated translations as preferred examples, with perturbed versions as rejections.
- **KTO-IT2-S.Gold**: IT2 generated translations as preferred examples, with perturbed gold-standard translations as rejections.

These configurations systematically evaluate the effectiveness of KTO in low resource translation, demonstrating its potential to outperform SFT under identical data constraints.

4.3 Training Configuration

We fine-tune the Llama 3.1 Instruct 8B model for 1 epoch for both the SFT and KTO experiments. Due to computational constraints, further experimentation with additional epochs or technique combinations was not feasible.

To enhance computational efficiency, we employ 4-bit quantization (Kim et al., 2024) using QLoRA (Dettmers et al., 2023) for parameter-efficient fine-tuning. The specific hyperparameter configurations for LoRA are outlined in Table 5.

5 Evaluation

We evaluated the translation quality of the fine-tuned models using BLEU[3] (Post, 2018) and CHRF[4] (Popović, 2015) on the dev and devtest splits of the Flores-200 Benchmark Dataset[5] (Costa-jussà et al., 2022). We used the sacreBLEU library for BLEU and chrF calculation. The Flores-200 dataset provided a comprehensive benchmark for evaluating machine translation across various language pairs.

[3] https://github.com/mjpost/BLEU
[4] https://github.com/marian-nmt/chrf
[5] https://github.com/facebookresearch/flores

6 Results and Discussion

6.1 Overall Performance Comparison (SFT vs. KTO)

Our experiments demonstrate the effectiveness of KTO-based preference alignment with progressive perturbation for low-resource Indian language translation. As shown in Figure 1, KTO consistently outperforms SFT in selected languages. In the Flores-200 devtest set, we observed an average BLEU improvement from 1.84 to 2.47 and CHRF from 2.85 to 4.01 compared to SFT. These gains, achieved with the same positive training data and computational constraints, highlight the data efficiency of our approach.

6.2 KTO Configuration Analysis

Among KTO variants, KTO_IT2_S.Gold achieved the highest scores, while KTO_Gold_S.IT2 performed the lowest.

Using IT2-generated translations as the preferred completion consistently outperformed gold-standard human translations, aligning with trends observed in SFT. This suggests that IT2 translations may provide a more effective learning signal than gold translations in our setup. Additionally, using perturbed gold translations (S.Gold) as rejected examples generally resulted in better model alignment than perturbed IT2 translations (S.IT2), likely due to the higher intrinsic quality of gold translations.

6.3 Language-Specific Observations

A notable exception was Santali, where SFT outperformed all KTO variants. This outcome is likely due to the model's limited initial proficiency in Santali. Since KTO relies on negative examples, it may amplify noise when the baseline quality is extremely low. In such cases, the model might learn

Model	Metric	English→XX					XX→English					Hin→XX			XX→Hin		
		Hin	Ban	Tam	Tel	Sat	Hin	Ban	Tam	Tel	Sat	Ban	Tam	Tel	Ban	Tam	Tel
Llama3.1-Instruct-8B	BLEU	17.43	7.64	4.54	7.43	0.03	32.71	25.14	19.79	24.92	0.63	7.51	4.03	6.12	14.16	9.36	11.52
	CHRF	45.04	39.33	38.74	37.44	2.40	60.74	54.30	49.44	53.41	18.88	38.88	37.68	36.66	39.88	34.50	36.64
Llama3.2-Instruct-3B	BLEU	7.97	4.04	4.10	6.01	0.01	10.14	9.07	8.38	10.56	0.03	6.11	4.14	5.33	4.61	3.45	5.30
	CHRF	36.75	35.35	39.88	38.66	2.06	47.49	44.56	42.27	46.16	3.47	37.25	37.90	35.99	27.39	25.75	30.66
Llama3.2-Instruct-11B	BLEU	17.06	7.08	4.46	6.60	0.01	30.94	25.29	19.95	26.24	1.63	7.31	3.95	5.71	13.87	8.69	12.22
	CHRF	44.68	38.23	40.22	37.76	2.65	59.85	54.62	50.19	54.59	23.41	38.70	38.01	36.19	40.51	33.42	38.41
SFT_Gold	BLEU	20.74	8.81	6.75	10.38	2.24	34.73	27.37	24.64	28.95	7.54	8.36	5.68	7.70	14.73	12.33	13.95
	CHRF	46.98	40.23	44.30	44.09	27.74	61.17	54.98	51.93	56.13	30.21	39.48	41.34	40.39	40.59	37.00	40.04
SFT_IT2	BLEU	21.05	10.15	8.20	11.10	**2.51**	35.25	28.30	25.45	30.15	**7.8**	9.10	6.25	7.35	15.30	12.20	14.15
	CHRF	48.20	43.05	45.35	46.25	**31.61**	61.05	55.30	52.45	56.15	**30.56**	41.15	43.20	42.35	42.10	38.25	40.10
KTO_Gold_S.IT2	BLEU	20.88	9.64	7.25	9.57	0.78	34.68	27.27	23.95	28.17	4.63	8.84	5.21	7.35	15.41	11.94	14.21
	CHRF	47.48	41.96	45.10	43.81	19.34	61.26	54.89	51.63	55.15	25.14	40.65	41.91	40.47	41.40	37.00	40.01
KTO_Gold_S.Gold	BLEU	21.15	9.44	7.04	10.10	0.70	34.31	27.57	24.28	28.19	4.50	8.50	5.67	7.18	14.74	12.31	13.89
	CHRF	48.04	42.05	44.82	44.08	19.22	61.13	55.40	51.76	55.49	25.18	40.69	41.50	40.30	40.94	37.33	39.96
KTO_IT2_S.IT2	BLEU	21.99	11.15	**8.71**	**11.49**	0.15	36.68	30.09	27.11	31.17	5.66	10.09	**7.32**	9.13	16.99	13.68	15.30
	CHRF	50.02	45.01	48.40	47.30	13.42	62.64	57.20	53.95	57.52	26.17	43.00	45.98	43.71	43.60	39.40	41.69
KTO_IT2_S.Gold	BLEU	**25.26**	**13.81**	6.00	10.09	0.43	**38.26**	**30.96**	**27.78**	**31.68**	7.46	**10.96**	6.35	**10.08**	**17.33**	**14.55**	**16.64**
	CHRF	**51.21**	**47.74**	**49.93**	**48.63**	14.56	**63.72**	**57.94**	**54.93**	**58.27**	28.18	**44.27**	**46.17**	**44.67**	**44.10**	**40.17**	**43.07**

Table 4: Performance comparison of Zero-Shot Llama models vs. SFT & KTO fine-tuned Llama 3.1 Instruct-8B on Flores DevTest. SFT models use supervised fine-tuning with either gold-standard human translations (SFT_Gold) or IndicTrans2-generated translations (SFT_IT2). KTO models apply Kahneman-Tversky Optimization with different preference and rejection criteria: gold-standard translations with perturbed IT2 (KTO_Gold_S.IT2), gold-standard translations with perturbed gold-standard translations (KTO_Gold_S.Gold), IT2 translations with perturbed IT2 (KTO_IT2_S.IT2), and IT2 translations with perturbed gold-standard translations (KTO_IT2_S.Gold). All SFT and KTO models are fine-tuned versions of Llama 3.1 Instruct-8B.

Method	**Value**
`LoRA modules`	PEFT
`Rank`	8
`Alpha`	8
`Dropout`	0
`Learning rate`	5e-5
`Effective batch size`	64
`Epochs`	1

Table 5: Hyper-parameter configurations for LoRA

to avoid all translation choices from the Santali data, including those that are correct.

7 Conclusion

This study explores KTO with progressive perturbation for Indian language translation, demonstrating its superiority over SFT in most cases and highlighting its potential to maximize the utility of existing data in resource-scarce scenarios. Our method systematically degrades high-quality translations through controlled perturbations, generating a spectrum of negative examples ranging from overtly erroneous to subtly flawed outputs. These negative samples provide a rich training signal, helping the model distinguish between accurate and error-prone translations, thereby enabling efficient learning from limited data.

Notably, IT2-generated translations were more effective than gold-standard translations as preferred completions, raising questions about the reliability of the gold data in the BPCC Dataset. However, KTO was less effective in extremely low-resource cases like Santali, where SFT outperformed it, suggesting that KTO's effectiveness depends on the model's initial proficiency in a given language.

8 Limitations

In conducting our experiments, we relied on high-performance GPUs, specifically RTXA6000. However, we acknowledge that not everyone may have access to such powerful computing resources, which could present challenges in reproducing our experiments and achieving identical results. Despite these computing limitations, we were still able to carry out meaningful experiments, although we were unable to conduct more comprehensive analyses.

9 Future Work

Future work could explore several directions, including experimenting with different perturbation schedules for performance improvements. Ad-

ditionally, addressing the challenges of applying KTO with progressive perturbation to low-resource languages like Santali is crucial, possibly by adapting the strategy or exploring alternative training objectives. Finally, applying this approach to other low-resource machine translation tasks across language families and domains could help assess its generalizability.

10 CO_2 Emission Related to Experiments

Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.813 $kgCO_2eq/kWh$. A cumulative of 648 hours of computation was performed on hardware of type RTX A6000 (TDP of 300W).

Total emissions are estimated to be 158.05 $kgCO_2eq$ of which 0 percent were directly offset.

Estimations were conducted using the Machine-Learning Impact calculator presented in (Lacoste et al., 2019).

References

Nishaant Choksi. 2018. Script as constellation among munda speakers: the case of santali. *South Asian History and Culture*, 9(1):92–115.

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*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

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.

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*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Hippolyte Gisserot-Boukhlef, Ricardo Rei, Emmanuel Malherbe, Céline Hudelot, Pierre Colombo, and Nuno M Guerreiro. 2024. Is preference alignment always the best option to enhance llm-based translation? an empirical analysis. *arXiv preprint arXiv:2409.20059*.

Vikrant Goyal, Sourav Kumar, and Dipti Misra Sharma. 2020. Efficient neural machine translation for low-resource languages via exploiting related languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168, Online. Association for Computational Linguistics.

Jeonghoon Kim, Jung Hyun Lee, Sungdong Kim, Joonsuk Park, Kang Min Yoo, Se Jung Kwon, and Dongsoo Lee. 2024. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization. *Advances in Neural Information Processing Systems*, 36.

Nidhi Kowtal, Tejas Deshpande, and Raviraj Joshi. 2024. A data selection approach for enhancing low resource machine translation using cross lingual sentence representations. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, pages 1–7. IEEE.

Arun Kumar, V Dhanalakshmi, RU Rekha, KP Soman, S Rajendran, et al. 2009. Morphological analyzer for agglutinative languages using machine learning approaches. In *2009 International Conference on Advances in Recent Technologies in Communication and Computing*, pages 433–435. IEEE.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020. Norm-based curriculum learning for neural machine translation. *Preprint*, arXiv:2006.02014.

Nick Mecklenburg, Yiyou Lin, Xiaoxiao Li, Daniel Holstein, Leonardo Nunes, Sara Malvar, Bruno Silva, Ranveer Chandra, Vijay Aski, Pavan Kumar Reddy Yannam, et al. 2024. Injecting new knowledge into large language models via supervised fine-tuning. *arXiv preprint arXiv:2404.00213*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom M. Mitchell. 2019. Competence-based curriculum learning for neural machine translation. *Preprint*, arXiv:1903.09848.

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, Belgium, Brussels. Association for Computational Linguistics.

Akshai Ramesh, Haque Usuf Uhana, Venkatesh Balavadhani Parthasarathy, Rejwanul Haque, and Andy Way. 2021. Augmenting training data for low-resource neural machine translation via bilingual word embeddings and bert language modelling. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Kengatharaiyer Sarveswaran, Gihan Dias, and Miriam Butt. 2021. Thamizhi morph: A morphological parser for the tamil language. *Machine Translation*, 35(1):37–70.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy. Association for Computational Linguistics.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. An empirical exploration of curriculum learning for neural machine translation. *Preprint*, arXiv:1811.00739.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

A Appendix

Model	Metric	English→XX					XX→English					Hin→XX			XX→Hin		
		Hin	Ban	Tam	Tel	Sat	Hin	Ban	Tam	Tel	Sat	Ban	Tam	Tel	Ban	Tam	Tel
Llama3.1-Instruct-8B	BLEU	17.37	7.45	4.63	7.21	0.01	31.06	26.31	20.85	26.60	0.53	7.56	4.32	6.00	14.18	10.15	12.01
	CHRF	45.76	39.54	39.05	37.89	2.16	60.23	55.12	50.37	54.82	18.35	39.33	38.29	36.75	39.95	35.39	37.20
Llama3.2-Instruct-3B	BLEU	7.61	2.93	3.64	5.97	0.01	8.79	8.11	7.99	10.90	0.02	7.24	4.13	5.45	4.29	3.02	4.98
	CHRF	36.64	34.11	38.49	39.31	1.96	45.34	43.49	41.42	46.84	3.53	39.12	37.79	36.14	27.30	24.26	30.97
Llama3.2-Instruct-11B	BLEU	17.30	8.04	4.68	6.14	0.04	31.36	25.49	21.24	27.17	1.42	6.83	3.31	5.76	14.09	9.66	12.65
	CHRF	45.15	40.39	40.91	36.78	3.05	60.03	55.00	50.70	55.36	22.83	38.87	34.70	36.69	41.20	34.72	39.24
SFT_Gold	BLEU	21.79	8.50	7.10	10.23	2.31	35.26	28.69	25.79	30.90	**7.62**	7.61	6.27	7.80	15.56	12.65	15.06
	CHRF	48.30	40.98	44.58	44.17	27.78	61.63	55.87	52.63	57.62	**30.18**	39.80	41.96	40.92	41.84	37.60	41.17
SFT_IT2	BLEU	23.45	12.34	**10.05**	12.30	**3.10**	36.15	30.25	27.40	32.10	6.05	9.15	7.25	8.40	17.20	14.35	16.50
	CHRF	50.12	45.67	48.24	47.10	**29.62**	62.34	57.12	54.30	58.45	25.10	43.15	45.30	43.20	44.05	41.25	43.40
KTO_Gold_S.IT2	BLEU	19.52	7.67	5.43	7.73	0.57	33.98	28.55	24.17	29.53	5.13	6.68	4.37	6.76	13.34	11.02	12.37
	CHRF	46.55	40.12	41.72	40.44	19.00	61.02	55.68	51.20	56.52	25.12	38.61	39.04	38.66	39.69	35.65	38.40
KTO_Gold_S.Gold	BLEU	22.05	9.71	7.69	10.35	0.96	33.98	28.55	24.91	29.60	5.15	8.18	6.02	7.36	16.00	12.62	14.83
	CHRF	48.87	42.81	45.38	44.20	19.23	61.02	55.68	51.73	56.61	25.26	41.10	42.04	40.56	42.62	37.51	41.15
KTO_IT2_S.IT2	BLEU	23.90	11.79	10.04	**12.33**	0.15	36.67	30.57	27.54	32.31	6.13	9.70	**7.83**	8.82	17.12	14.66	16.53
	CHRF	51.16	45.98	49.32	47.13	12.87	62.59	57.86	54.56	58.53	26.21	43.49	46.16	43.60	44.24	40.56	42.85
KTO_IT2_S.Gold	BLEU	**26.37**	**14.01**	5.60	12.01	0.34	**38.78**	**32.14**	**28.56**	**33.90**	7.33	**10.93**	6.77	**10.30**	**18.24**	**15.60**	**16.98**
	CHRF	**52.08**	**48.55**	**49.44**	**49.15**	14.02	**63.84**	**59.07**	**55.26**	**60.00**	28.18	**44.69**	**46.57**	**45.12**	**45.07**	**40.96**	**43.53**

Table 6: Performance comparison of Zero-Shot Llama models vs. SFT & KTO fine-tuned Llama 3.1 Instruct-8B on Flores Dev. SFT models use supervised fine-tuning with either gold-standard human translations (SFT_Gold) or IndicTrans2-generated translations (SFT_IT2). KTO models apply Kahneman-Tversky Optimization with different preference and rejection criteria: gold-standard translations with perturbed IT2 (KTO_Gold_S.IT2), gold-standard translations with perturbed gold-standard translations (KTO_Gold_S.Gold), IT2 translations with perturbed IT2 (KTO_IT2_S.IT2), and IT2 translations with perturbed gold-standard translations (KTO_IT2_S.Gold). All SFT and KTO models are fine-tuned versions of Llama 3.1 Instruct-8B.

Leveraging Visual Scene Graph to Enhance Translation Quality in Multimodal Machine Translation

Ali Hatami[1], Mihael Arcan[2], Paul Buitelaar[1]

[1]Insight Research Ireland Centre for Data Analytics,
Data Science Institute, University of Galway, Ireland
[2]Lua Health, Galway, Ireland,

Correspondence: ali.hatami@insight-centre.org

Abstract

Despite significant advancements in Multimodal Machine Translation, understanding and effectively utilising visual scenes within multimodal models remains a complex challenge. Extracting comprehensive and relevant visual features requires extensive and detailed input data to ensure the model accurately captures objects, their attributes, and relationships within a scene. In this paper, we explore using visual scene graphs extracted from images to enhance the performance of translation models. We investigate this approach for integrating Visual Scene Graph information into translation models, focusing on representing this information in a semantic structure rather than relying on raw image data. The performance of our approach was evaluated on the Multi30K dataset for English into German, French, and Czech translations using BLEU, chrF2, TER and COMET metrics. Our results demonstrate that utilising visual scene graph information improves translation performance. Using information on semantic structure can improve the multimodal baseline model, leading to better contextual understanding and translation accuracy.

1 Introduction

Neural Machine Translation (NMT) has significantly advanced translation quality compared to earlier methods, showcasing remarkable improvements in fluency and precision (Cho et al., 2014). Transformer-based models enhanced performance by effectively capturing semantic dependencies and producing fluent, contextually relevant translations (Vaswani et al., 2017).

However, despite these advancements, text-only NMT models face persistent challenges in translating the input text (Wang and Xiong, 2021; Zhao et al., 2022). Resolving ambiguity in the input sentence is one of these challenges (Futeral et al., 2023; Bowen et al., 2024; Hatami et al., 2024).

To address these limitations, researchers have explored Multimodal Machine Translation (MMT), a subfield of NMT that integrates visual information from images or videos to enhance translation models (Yao and Wan, 2020; Wang and Xiong, 2021; Zhao et al., 2022). MMT leverages visual content as a complementary source of information to aid in understanding the source text and resolving ambiguities. Text-only NMT models might struggle to translate ambiguous sentences, but an accompanying image can provide crucial visual cues for disambiguation, enabling the model to select the correct translation.

Despite its potential, MMT presents its own challenges. Visual resources, such as images, often contain a large amount of information, not all of which is relevant to the translation task. This extra information can not only fail to improve translation quality but may even degrade it. In addition, training an MMT model requires a vast amount of visual information covering different objects and their relationships.

To address these challenges, recent studies have focused on identifying and incorporating the most relevant visual information into translation models (Lala and Specia, 2018; Fei et al., 2023; Yin et al., 2023; Hatami et al., 2023). These papers examine the importance of using visual information by focusing on lexical ambiguity in the input text to find relevant information on the visual side.

In this paper, we study the impact of using Visual Scene Graphs (VSGs), which represent objects and their relationships within an image, as a means to enhance MMT models. First, we extract VSGs as a semantic structure from images and then utilize this information as triples to train our translation model. Our work differs from previous studies by directly leveraging VSGs to represent objects and their relationships, providing a structured se-

Proceedings of Machine Translation Summit XX Volume 1, pages 353–364, June 23-27, 2025

mantic context for translation. We evaluated our approach on the Multi30K dataset for English into German, French and Czech translations. The results demonstrate that the use of VSGs in MMT leads to notable improvements in both quantitative metrics and qualitative evaluations, highlighting the potential of this approach for advancing the field of multimodal translation.

2 Related Work

In recent years, MMT has gained significant attention to enhance traditional text-only translation by incorporating visual information. MMT models primarily relied on image features extracted from vision-based transformers to improve translation quality, particularly in cases of ambiguity or lexical uncertainty (Delbrouck and Dupont, 2017). Early approaches to MMT incorporated joint multimodal embeddings to fuse textual and visual features. Calixto et al. (2017) proposed an attention-based framework that used convolutional neural networks (CNNs) to extract image features, which were then integrated into a sequence-to-sequence NMT model. Similarly, Libovický and Helcl (2017) introduced hierarchical attention mechanisms to balance contributions from different modalities dynamically.

Some other papers explored transformer-based architectures to enhance multimodal fusion. Wu et al. (2021) adapted the Transformer model by introducing multimodal self-attention, enabling better integration of visual and textual features. Caglayan et al. (2019) demonstrated that incorporating region-based visual features (e.g., using object detectors like Faster R-CNN) improved MMT performance by focusing on semantically relevant image regions.

Despite advancements, challenges remain in effectively integrating multimodal information without introducing noise. Elliott (2018) found that while images help in specific cases, text-only models often outperform multimodal ones when trained on large-scale datasets. This has led to investigations into adaptive multimodal fusion techniques, where the model selectively uses visual information only when beneficial (Hatami et al., 2024).

Recent advancements in MMT have explored the integration of structured visual knowledge to enhance translation quality. Yin et al. (2020) proposed a graph-based multimodal fusion encoder for NMT, leveraging Graph Neural Networks (GNNs) to encode multimodal information more effectively. By structuring both visual and textual inputs into a graph representation, their model captures semantic relationships between objects, improving the contextual grounding of translations. These studies highlight the growing importance of structured vision-language representations, such as scene graphs and graph-based encoders, in addressing the challenges of multimodal translation, particularly in ambiguous and resource-constrained settings.

Incorporating knowledge graphs into NMT has proven effective in improving the translation of named entities and specialized terminology, as demonstrated by Moussallem et al. (2019). Their approach introduced two strategies: Entity Linking with Knowledge Bases, which enriched NMT embeddings through multilingual entity linking, and Surface Form Initialization, which optimized entity vector values without explicit linking. By leveraging structured knowledge representations, their method enhanced translation accuracy, particularly in handling domain-specific terms and low-resource scenarios.

Unsupervised MMT (UMMT) system introduced by Fei et al. (2023) that utilises scene graphs as a pivoting mechanism to perform inference-time image-free translation through visual scene hallucination. Their method generates synthetic scene graphs from textual input, enabling multimodal translation even in the absence of actual image inputs. This approach effectively bridges the gap between vision and language representations, demonstrating improved translation performance in low-resource and zero-resource scenarios.

Although VSGs are widely used in various multimodal tasks such as image captioning (Yang et al., 2018), visual question answering (Hildebrandt et al., 2020), and image retrieval (Johnson et al., 2018), they remain underexplored in the multimodal translation task. VSGs provide a powerful representation for understanding image semantics by capturing objects, their attributes, and relationships in a structured graph format. In the context of MMT, leveraging the structured and interpretable visual information provided by scene graphs has the potential to enhance the translation process by improving contextual grounding and disambiguating visually dependent terms.

In our work, we propose an approach by leveraging VSGs extracted using a Multimodal Large Language Model (MLLM) to improve translation quality in MMT systems. By using MLLMs, we

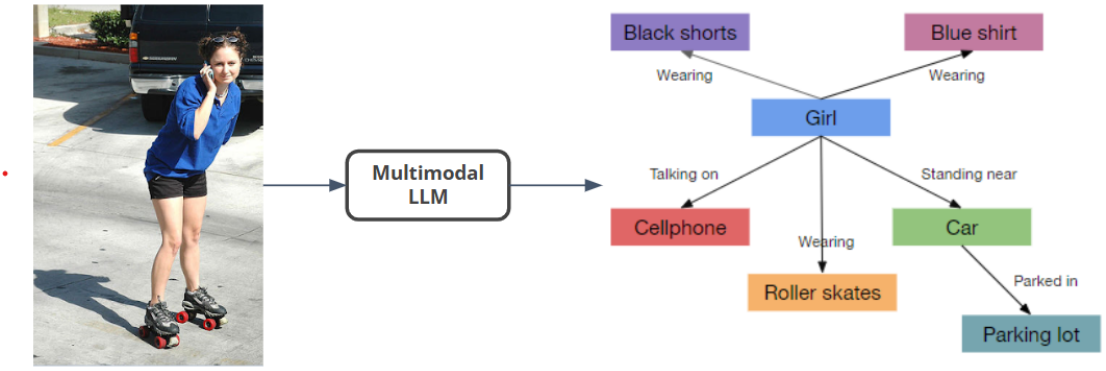


Figure 1: Example for extracting a Visual Scene Graph (VSG) from an image.

ensure accurate and detailed scene graph extraction, capturing not only objects and their relationships but also contextual nuances often missed by conventional visual models. This structured visual information is then incorporated into the translation pipeline, enabling our model to produce translations that are more contextually appropriate and semantically accurate. Figure 1 shows an example of the VSGs extracted from an image using Gemini 1.5 Flash.

To the best of our knowledge, few studies focus on extracting object relationships in MMT (Fei et al., 2023; Yin et al., 2023). By integrating scene-graph information into translation models, we aim to address the limitations of raw visual inputs and provide meaningful context for disambiguation and improved translations. Unlike prior approaches that focus on multimodal fusion without explicit scene-graph extraction or rely on hallucinated visual representations during inference, we extract VSGs from images and utilize them as triples to enhance translation quality through structured semantic learning. The integration of triples aims to provide contextual information about the scene, potentially disambiguating lexical or syntactic ambiguities in the text. Our results demonstrate that incorporating the VSG information yields better performance compared to using raw images as visual input.

3 Methodology

In this section, we explain our methodology for extracting scene graph information from images and utilising it in the translation process.

3.1 Visual Scene Graph Extraction

To integrate visual information into the translation model, we extract Visual Scene Graphs (VSGs) in English from images. VSGs provide structured representations of images in a triple format (subject, relationship, object), capturing object relationships and semantic context. This structure encodes visual information in a textual format, covering all objects and their relationships within the scene.

Prompt

```
Extract the visual scene graph as triples from the
provided image and save it in a Python list.

Number identical objects if more than one exists,
and ensure the visual scene graph is in English.
```

```
[("girl", "wearing", "black shorts"), ("girl",
"wearing", "blue shirt"), ("girl", "talking
on", "cellphone"), ("girl", "wearing", "roller
skates"), ("girl", "standing near", "car"),
("car", "parked in", "parking lot")]
```

Figure 2: Prompt example for extracting a Visual Scene Graph (VSG) from an image in triples format using Gemini.

We use Gemini 1.5 Flash as a multimodal LLM to generate Visual Scene Graphs (VSGs) from images. Gemini includes parameters such as temperature, top_P, and safety settings to control generating the output. These parameters are explained in Section 4.2 in more detail. After configuring these parameters, the model generates VSGs from images for the training, validation, and test sets based on the provided prompt. Figure 2 shows the prompt used to extract VSG from the given image.

To ensure a consistent output format, we enforced the model to generate VSGs in a Python list, preventing variations in format. We also restricted the model to generate VSGs strictly in English to reduce hallucinations, as it sometimes defaulted to other languages based on the image context. Ad-

Prompt 1

```
Translate the following English sentence to German:

A trendy girl talking on her cellphone while gliding
slowly down the street.
```

Prompt 2

```
Translate the following English sentence to German:

A trendy girl talking on her cellphone while gliding
slowly down the street.

Use the following triples to ensure the translation
is correct:

girl | wearing | black shorts
girl | wearing | blue shirt
girl | talking on | cellphone
girl | wearing | roller skates
girl | standing near | car
car | parked in | parking lot
```

Prompt 4

```
Translate the following English sentence to German:

A trendy girl talking on her cellphone while gliding
slowly down the street.

Use the following triples and image to ensure the
translation is correct:

girl | wearing | black shorts
girl | wearing | blue shirt
girl | talking on | cellphone
girl | wearing | roller skates
girl | standing near | car
car | parked in | parking lot
```

Prompt 3

```
Translate the following English sentence to German:

A trendy girl talking on her cellphone while gliding
slowly down the street.

Use the following image to ensure the translation is
correct:
```

Figure 3: Prompt examples that we used for T5 and Gemini to translate the input text from English to German; Prompt 1: Text-to-Text translation, Prompt 2: Text+Triples-to-Text translation, Prompt 3: Text+Image-to-Text translation, Prompt 4: Text+Triples+Image-to-Text translation.

ditionally, we numbered identical objects in the VSGs to improve scene comprehension when multiple identical objects were present.

3.2 Training Text-to-Text Model

Text-to-Text (T2T) translation is a baseline approach in which the model is used to translate the input text from the source language into the target language. For T2T translation, we utilise four models: NMT-T2T, mT5_Base, NLLB-200, and Gemini. NMT-T2T is a transformer-based model trained on the dataset, while mT5_Base and NLLB-200 are fine-tuned on the dataset. Additionally, we use Gemini for zero-shot translation of the test sets.

Prompt 1 in Figure 3 illustrates an example prompt used for mT5 and Gemini to translate the input sentence from English into German. Unlike mT5 and Gemini, which are multitask models requiring prompt instructions for translation, NLLB-200 is specifically trained for translation tasks. Therefore, we simply provide the input sentence to the fine-tuned NLLB-200 model to generate the translation.

3.3 Training Text+Triplets-to-Text Model

To investigate the impact of incorporating VSG, we enriched the source text with the information extracted from VSG. In Text+Triples-to-Text (TT2T) translation, we incorporate this information (Section 3.1) into the training process of the translation model. By augmenting the text with structured visual-contextual information, we aimed to assess whether the inclusion of triples improves the ability of the models to capture implicit meanings and context that are otherwise absent in text-only inputs.

For TT2T translation, we concatenate these triples with the English input text to provide additional context, helping the model better understand the input. This approach leverages semantic insights from visual relationships in a textual format, enhancing translation quality without directly using images. Similar to T2T, in TT2T, we utilise four models: NMT-T2T, mT5_Base, NLLB-200, and Gemini. We train NMT-T2T on input text enriched with triple information, along with the corresponding output text. We also fine-tune mT5_Base and NLLB-200 on input text enriched with triple information. For Gemini, we apply zero-shot translation to translate test set sentences while incorporating triple information to ensure accurate translation.

Prompt 2 in Figure 3 presents an example prompt used for mT5 and Gemini to translate an input sentence from English to German. By adding triples extracted from the paired image, we guide the model to consider semantic information from the image when translating. This approach ensures

that the translation aligns correctly with the visual context.

3.4 Training Text+Image-to-Text Model

In Text+Image-to-Text (TI2T) translation, we use the input text along with an image to train the model. For TI2T translation, we utilise two models: MMT-TI2T and Gemini. MMT-TI2T is a gated fusion multimodal model trained on the training and validation sets. For Gemini, we use zero-shot translation of the given sentence, considering the paired image. Prompt 3 in Figure 3 indicates the example prompt in TT2T translation from English to German. In the prompt, we provide an instruction to the model to use the given image to make sure the translation is correct.

3.5 Training Text+Triplets+Image-to-Text Model

For Text+Triples+Image-to-Text (TTI2T) translation, we add triples extracted from Visual Scene Graphs (VSGs) as additional information to the translation model alongside the input text and image. The reason behind this approach is that using images alone may introduce noise and degrade the performance of the translation model. By incorporating structured semantic information from the scene graph along with the image, enables the model to incorporate both low-level visual details and high-level relational knowledge into the translation process.

For TTI2T, we employ two multimodal translation models: MMT-TI2T and Gemini. We explain both models in Section 3.4. The only difference is that TTI2T additionally provides extracted triples along with the input text and image.

Prompt 4 in Figure 3 shows an example prompt for TTI2T translation from English to German. In the prompt, we instruct the model to use the given image and triples to ensure the translation is accurate.

4 Experimental Setup

In this section, we provide insights into the dataset used in this work, extracting VSG from images, settings for text-only and multimodal models, and the translation evaluation metrics BLEU, ChrF2, TER and COMET.

4.1 Multi30k Dataset

Multi30K (Elliott et al., 2016) is an extension of the Flickr30K Entities dataset that consists of 29,000 images paired with descriptions in English, along with translated sentences in German, French, and Czech (Elliott et al., 2017). The dataset is specifically designed for evaluating MMT systems, where both textual and visual information are utilised for translation tasks. Multi30K also provides three test sets: the 2016 and 2017 test sets, each with 1,000 images, and the 2018 test set with 1,071 images.

4.2 Gemini 1.5 Flash

To extract VSGs from the Multi30K dataset, we used Gemini 1.5 Flash [1], a pre-trained LLM to analyse the multimodal data. For our experiment, we used Gemini through the free-tier API, which provides a rate limit of 15 requests per minute (RPM) and 1,500 requests per day (RPD). We set the default inference parameters for the model. These defaults included a temperature of 1.0, ensuring a balanced mix of randomness and determinism in responses, a Top-p sampling set to 0.95, allowing diverse but high-probability token selections, and a maximum output length of 8,192 tokens. The default Top-k setting was automatically adjusted by the system. To ensure comprehensive processing of all images in the dataset, we configured the model's safety settings, including thresholds for "Harassment", "Hate Speech", "Sexually Explicit Content", and "Dangerous Content" to "BLOCK_NONE". This adjustment allows the model to generate responses for every image ensuring that outputs are returned in full without being restricted by safety mechanisms. After setting the parameters, the model generated VSG from the image in our dataset based on the given prompt (Figure 2).

Gemini 1.5 Flash is capable of processing both text and visual information. For text-only and multimodal translation, we also employed Gemini, maintaining the same parameter settings and safety configurations as described in VSG extraction. The model was used for zero-shot translation from English into German, French, and Czech on the Multi30k dataset, covering both text-only and multimodal translation under different configurations. These configurations included T2T (En $\rightarrow$ De, Fr, Cs), TT2T (En + triples $\rightarrow$ De, Fr, Cs), TI2T (En + image $\rightarrow$ De, Fr, Cs), and TIT2T (En + image + triples $\rightarrow$ De, Fr, Cs). This setup allowed us to assess Gemini's capability in handling both textual and multimodal inputs across multiple

[1] https://deepmind.google/technologies/gemini/

languages.

4.3 OpenNMT

A text-only transformer model serves as the baseline in our experiment, utilising solely the textual captions of images for translation. Trained using the OpenNMT toolkit (Klein et al., 2018) on the Multi30k dataset for English to German, French, and Czech translations, the model comprises a 6-layer transformer architecture with attention mechanisms in both encoder and decoder stages, trained for 50K steps. Sentencepiece (Kudo and Richardson, 2018) is employed to segment words into subword units, offering a language-independent approach to tokenization without necessitating preprocessing steps, thus enhancing the model's adaptability and versatility in handling raw text.

4.4 Gated Fusion Multimodal

In the MMT model, we adopt the gated fusion MMT model (Wu et al., 2021) as a multimodal basline model. Gated fusion is a mechanism that is used to integrate visual information from images with textual information from source sentences by fusing visual and text representations by employing a gate mechanism.. The main idea behind gated fusion is to control the amount of visual information that is blended into the textual representation using a gating matrix. The source sentence x is fed into a vanilla Transformer encoder to obtain a textual representation H_{text} of dimension $T \times d$. The image z is processed using a pre-trained ResNet-50 CNN which has been trained on the ImageNet dataset (Deng et al., 2009) to extract a 2048-dimensional average-pooled visual representation, denoted as $Embed_{image}(z)$. The visual representation $Embed_{image}(z)$ is projected to the same dimension as H_{text} using a weight matrix W_z. A gating matrix Λ of dimension $T \times d$ is generated to control the fusion of the textual and visual representations. The gating matrix Λ is computed as:

$$\Lambda = \text{sigmoid}(W_\Lambda \text{Embed}_{\text{image}}(z) + U_\Lambda H_{\text{text}})$$

where W_Λ and U_Λ are model parameters.

4.5 NLLB-200

In this section, we outline the setup used the No Language Left Behind (NLLB) model. This model is a transformer-based multilingual NMT model designed for covering 200 languages. Due to our GPU limitation, we fine-tune NLLB-200 with 600M model on our dataset. The process involved data preprocessing, model training, hyperparameter tuning, and evaluation.

Similar to mT5, the fine-tuning process was conducted using two NVIDIA A6000 GPUs (2×48GB GPU memory). We set the learning rate to 2e-5 and used the Adam optimizer with a weight decay of 0.01 to prevent overfitting. The model was trained for 10 epochs with a per-device batch size of 16 for both training and evaluation. To ensure efficient monitoring, logging was performed every 500 steps. The training leveraged Automatic Mixed Precision (AMP) for optimized memory usage and performance.

4.6 Multilingual T5

Multilingual Text-to-Text Transfer Transformer (mT5) is a transformer-based language model designed specifically for multilingual Natural Language Processing (NLP) tasks. It extends the T5 model, which frames all NLP tasks as text-to-text problems (Raffel et al., 2020). We fine-tuned the mT5 model on the Multi30K dataset to optimise its performance in translation tasks, focusing solely on the textual modality without any information from the visual side.

One of the key features of mT5 is its support for 101 languages, making it a powerful model for multilingual applications such as translation tasks (Xue et al., 2021). The model is pretrained on mC4 (Multilingual Common Crawl), a large-scale dataset containing filtered web text from a wide range of languages. This extensive training allows mT5 to perform well in both high-resource and low-resource languages. Additionally, since mT5 is trained on a diverse dataset, it is more capable of handling syntactic and grammatical variations across different languages (Raffel et al., 2020). Supporting multiple languages makes it well-suited for machine translation, allowing us to leverage a single model without the need for separate models for different languages.

We used mT5-Base which has around 220 M parameters. When fine-tuning mT5, common settings include a learning rate of 2e-5, which helps to ensure stable convergence during training while avoiding overfitting. The batch size is set to 16 for both training and evaluation, which balances efficiency and memory constraints, though it can be adjusted depending on GPU availability. Additionally, a weight decay of 0.01 is used to reduce

	English → German				English → French				English → Czech			
	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑
Text-to-Text (T2T)												
NMT-T2T	41.1	65.4	43.8	0.8604	60.6	71.4	31.8	0.8765	31.8	56.4	49.8	0.8852
mT5_Base	36.8	62.1	46.7	0.8072	52.7	70.5	32.4	0.8255	27.4	50.7	54.5	0.8109
NLLB-200	44.0*†	68.7*†	41.2*	0.862	66.4*†	80.3*†	22.3*†	**0.8916**	**37.6***†	**61.3***†	**44.7***†	0.8867
Gemini 1.5 Flash	43.7*†	68.7*†	41.2*	0.8657	54.5	73.2*	30.9	0.8755	35.0*†	59.9*	47.4*	0.8929
Text+Triplets-to-Text (TT2T)												
NMT-TT2T	41.3	65.7	43.6	0.8618	60.5	71.3	31.6	0.8779	31.9	56.6	49.7	0.8854
mT5_Base	37.2	62.5	46.0	0.8107	52.7	70.5	32.8	0.8266	27.7	51.1	54.4	0.8167
NLLB-200	44.6*†	69.1*†	40.7*†	0.8626	**67.0***†	**80.5***†	**21.9***†	0.8912	36.9*†	60.7*†	45.5*†	0.8828
Gemini 1.5 Flash	43.9*	68.7*†	40.8*†	0.8688	54.5	73.2	30.6	0.8803	34.5*†	59.2*	48.0	0.8923
Text+Image-to-Text (TI2T)												
MMT-TI2T	42.3*	66.6*	42.1*	0.8672	62.1*	72.6	31.1	0.8786	32.7	58.2*	47.6*	0.8864
Gemini 1.5 Flash	44.1*†	68.7*†	40.3*†	0.868	55.0	73.5*	30.8	0.8738	35.0*†	59.7*	48.4	0.8917
Text+Triplets+Image-to-Text (TTI2T)												
MMT-TTI2T	42.6*	66.8*	41.8*	0.8681	62.2*	72.5	30.9	0.8791	32.9	58.1*	47.8*	0.8862
Gemini 1.5 Flash	**45.1***†	**69.2***†	**40.1***†	**0.8696**	54.6	73.5*	30.4*	0.8767	34.8*†	59.7*	48.3	**0.8964**

Table 1: BLEU, ChrF2, TER and COMET scores for baseline and proposed models for English to German, French and Czech on the 2016 test set ($*$ and † represent a statistically significant results compared to baseline NMT and MMT respectively at a significance level of $p < 0.05$).

the risk of overfitting by penalizing excessively large model weights. We fine-tuned the model for 10 epochs by monitoring the validation loss during training to prevent unnecessary computations and potential overfitting. During training, logging every 500 steps provides periodic updates on performance, ensuring that any issues can be quickly identified and addressed.

4.7 Evaluation Metrics

We use four evaluation metrics: BLEU (Papineni et al., 2002), ChrF2 (Popović, 2015), TER (Snover et al., 2006), and COMET (Rei et al., 2020). BLEU assesses translation precision by comparing candidate translations to reference translations based on *n-grams*. ChrF2 evaluates the similarity between character *n-grams* in machine-generated and reference translations, particularly beneficial for languages with complex writing systems. TER quantifies the number of edits needed to align machine translations with human-generated references. COMET [2] is a neural-based metric that leverages both source and reference sentences to produce quality assessments aligned with human judgments. We conduct statistical significance testing using the *sacrebleu*[3] toolbox.

5 Results

In this section, we present the results of different translation models for language pairs of English into German, French and Czech. The evaluation is based on four metrics: BLEU, ChrF2, TER and COMET. In the first part, we focus on quantitative analysis, and in the second part, we conduct a qualitative analysis to manually evaluate the translation outputs of the models.

5.1 Quantitative Analysis

Table 1 presents the evaluation scores for our proposed multimodal and text-only translation models across English to German, French, and Czech translation tasks for the 2016 test set from the Multi30k dataset. For English to German translation, the Gemini (TTI2T) model achieved the highest scores in BLEU (45.1), ChrF2 (69.2), and COMET (0.8696) while also maintaining the lowest TER (40.1). This indicates that the inclusion of both triples and images in the input significantly enhanced translation quality. The NLLB-200 (TT2T) model closely followed, showing competitive results, particularly in ChrF2 (69.1) and COMET (0.8626). This suggests that leveraging structured data, even without images, is beneficial. Meanwhile, for English to French, the NLLB-200 (TT2T) model outperformed others with the highest BLEU (67.0) and lowest TER (21.9), showcasing its efficiency in maintaining fluency and adequacy. However, Gemini (TTI2T) scored the highest in COMET (0.8767), indicating that it produced the most human-like translations despite slightly lower BLEU. For English to Czech, NLLB-200 (T2T) led in all metrics, except COMET, where Gemini (TI2T) achieved the highest score (0.8929), emphasizing the benefit of incorporating multimodal

[2] https://github.com/Unbabel/COMET
[3] https://github.com/mjpost/sacrebleu

	English → German				English → French			
	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑
Text-to-Text (T2T)								
NMT-T2T	35.4	61.7	51.3	0.8548	49.4	68.6	35.8	0.8761
mT5_Base	29.9	57.3	55.8	0.7829	45.3	65.7	38.4	0.8169
NLLB-200	39.4*†	66.5*†	46.4*†	0.8566	**59.9***†	**76.8***†	**26.8***†	**0.8839**
Gemini 1.5 Flash	40.0*†	66.2*†	46.4*†	0.8632	53.1*†	73.2*†	32.0*†	0.8804
Text+Triplets-to-Text (TT2T)								
NMT-TT2T	35.3	61.5	51.6	0.8554	49.5	68.5	36.1	0.8723
mT5_Base	29.8	57.4	55.9	0.7796	45.5	65.7	38.8	0.8134
NLLB-200	38.1*†	65.7*†	48.9*	0.8504	59.5*†	76.4*†	27.9*†	0.8815
Gemini 1.5 Flash	39.8*†	66.2*†	45.8*†	0.863	52.5*	72.7*	32.5*	0.8737
Text+Image-to-Text (TI2T)								
MMT-TI2T	36.8	62.8	49.4	0.8572	51.3	71.5*	33.7	0.8768
Gemini 1.5 Flash	39.9*†	66.3*†	46.2*†	0.8624	54.3*†	73.6*†	31.7*	0.8786
Text+Triplets+Image-to-Text (TTI2T)								
MMT-TTI2T	37.1*	63.3	48.5*	0.8586	51.5	71.4	33.6	0.8781
Gemini 1.5 Flash	**40.6***†	**66.9***†	**45.4***†	**0.865**	53.9*†	73.6*†	31.5*†	0.8814

Table 2: BLEU, ChrF2, TER and COMET scores for baseline and proposed models for English to German and French on the 2017 test set ($*$ and † represent a statistically significant results compared to baseline NMT and MMT respectively at a significance level of $p < 0.05$).

information.

Gemini (TTI2T) consistently achieved top-tier scores, highlighting the advantages of integrating text, triples, and images across all language pairs. The lower BLEU and higher TER for mT5_Base across the board suggest its weaker ability to capture linguistic nuances. Notably, models using additional structured data (TT2T and TI2T) generally performed better than pure text-only models, confirming the effectiveness of multimodal approaches.

Table 2 presents the evaluation scores for our proposed multimodal and text-only translation models across English to German and French translation tasks for the 2016 test set from the Multi30k dataset. For English to German, Gemini (TTI2T) achieved the highest BLEU (40.6), ChrF2 (66.9), and COMET (0.865), along with the lowest TER (45.4). This again confirms the model's ability to leverage triplets and images to improve translation quality. Interestingly, NLLB-200 (T2T) performed best among text-only models, demonstrating its robustness. For English to French, NLLB-200 (T2T) set the highest scores in BLEU (59.9), ChrF2 (76.8), and TER (26.8), suggesting that its architecture excels in handling sentence-level fluency. However, Gemini (TTI2T) achieved the highest COMET (0.8814), implying that its translations were more aligned with human preferences.

Across both language pairs, Gemini (TTI2T) and NLLB-200 (T2T) consistently dominated, with the former benefiting from multimodal inputs and the latter excelling in text-based scenarios. Compared to 2016, TER values increased slightly, indicating a possible complexity shift in the test data. Overall, the performance gaps between text-only and multimodal models further widened, reinforcing the importance of multimodal approaches.

Table 3 presents the evaluation scores for our proposed multimodal and text-only translation models across English to German, French, and Czech translation tasks for the 2016 test set from the Multi30k dataset. For English to German, Gemini (T2T) outperformed all models in BLEU (37.6), TER (49.9), and COMET (0.8519), while Gemini (TI2T) led in ChrF2 (64.0). This suggests that including images provides more lexical coverage, enhancing character-level similarity. In English to French, NLLB-200 (TT2T) obtained the highest BLEU (43.1), while Gemini (TTI2T) dominated COMET (0.8503) and had the lowest TER (40.9), reinforcing the effectiveness of triples-based multimodal training. For English to Czech, NLLB-200 (TT2T) showed the highest BLEU (34.7), but Gemini (TTI2T) again achieved the highest COMET (0.8882), demonstrating improved translation quality with respect to human preferences.

Compared to 2016 and 2017, BLEU scores declined slightly in 2018, suggesting that the 2018 test set was more challenging. However, models incorporating multimodal inputs consistently performed better, emphasizing their enhanced ability to handle complex translation tasks. The consistently strong COMET scores achieved by Gemini

	English → German				English → French				English → Czech			
	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑	BLEU ↑	ChrF2 ↑	TER ↓	COMET ↑
Text-to-Text (T2T)												
NMT-T2T	32.4	59.8	54.6	0.8352	38.9	62.7	45.5	0.8418	28.9	52.8	57.4	0.8663
mT5_Base	28.1	55.2	58.9	0.7656	34.1	58.3	48.8	0.778	21.8	46.2	62.6	0.757
NLLB-200	37.3*†	63.5*	50.5*	0.8365	42.8*†	65.7*†	**40.8***†	0.8429	34.4*†	59.2*†	**49.9***†	0.8688
Gemini 1.5 Flash	**37.6***†	63.9*	**49.9***†	**0.8519**	42.3*†	65.6*	41.5*†	0.8475	33.2*†	**59.4***†	51.5*†	0.8877
Text+Triplets-to-Text (TT2T)												
NMT-TT2T	32.2	59.4	54.9	0.8346	39.1	62.8	45.5	0.8407	28.8	52.8	57.2	0.8641
mT5_Base	28.4	55.4	59.2	0.7678	34.3	58.4	48.9	0.7806	22.1	46.5	61.8	0.7628
NLLB-200	37.0*†	63.4*	51.3*	0.8351	**43.1***†	**65.8***†	41.1*†	0.8414	**34.7***†	59.2*†	50.8*†	0.8672
Gemini 1.5 Flash	37.0*†	63.7*	50.2*†	0.85	41.0	64.6*	42.3*	0.844	32.6*	58.5*†	51.8*†	0.8852
Text+Image-to-Text (TI2T)												
MMT-TI2T	33.7	61.2	52.4	0.8364	39.9	63.6	43.8	0.8485	30.1	54.8*	55.4*	0.8687
Gemini 1.5 Flash	37.0*†	**64.0***	50.4*	0.8506	42.4*	65.5*	41.3*	0.8476	33.1*†	58.7*†	52.2*†	0.8851
Text+Triplets+Image-to-Text (TTI2T)												
MMT-TTI2T	33.6	61.3	52.6*	0.8385	40.1	63.4	43.5*	0.847	30.3	54.7*	55.3*	0.8664
Gemini 1.5 Flash	37.2*†	63.3*	50.3*†	**0.8519**	42.6*	65.7*	40.9*†	**0.8503**	32.7*	58.5*†	52.7*†	**0.8882**

Table 3: BLEU, ChrF2, TER and COMET scores for baseline and proposed models for English to German, French and Czech on the 2018 test set ($*$ and † represent a statistically significant results compared to baseline NMT and MMT respectively at a significance level of $p < 0.05$).

(TTI2T) across all language pairs further underline its potential to produce translations that align more closely with human judgments.

Across the three test sets, the best-performing models varied depending on the language pair and evaluation metric. For English to German translation, the Gemini model showed the most significant improvement, particularly in the TTI2T setting. In English to French, the NLLB-200 model consistently outperformed others, especially in T2T translation. For English to Czech, the same model demonstrated strong performance. Overall, the results indicate that incorporating multimodal data, such as images and structured triples, enhances translation quality, with the TTI2T setting often achieving the best performance. These findings suggest that advanced multimodal approaches, particularly leveraging large-scale models like Gemini, can efficiently benefit from multimodal information and significantly improve machine translation across multiple languages and evaluation benchmarks.

5.2 Qualitative Analysis

In this section, we present examples from translation outputs to qualitatively analyse the performance of the models. We calculated sentence-level BLEU scores for each translation model and manually compared the translation quality across all sentences. Figure 4 shows two examples from the 2016 test set of the Multi30K data set: one for English to German and one for English to French translation.

In English to German, Gemini (TTI2T) provides the most accurate translation as it is identical to the reference sentence. This indicates that it perfectly preserves the original sentence's word choice, structure, and meaning. Specifically, it correctly translates "A boy wearing a red shirt" as "Ein Junge in einem roten Shirt", maintaining both the phrasing and natural German expression. Gemini (TI2T) is slightly less accurate but still acceptable. The only difference is the phrase "mit rotem Shirt" instead of "in einem roten Shirt." While both are grammatically correct, "in einem roten Shirt" is the more natural way to describe someone wearing a shirt in German. NLLB-200 (T2T) produces the weakest translation compared to Gemini. It translates "red shirt" as "roten Hemd," where "Hemd" usually refers to a button-down shirt rather than the more general "Shirt" in English. Also, NLLB-200 translates "into the sand" as "in den Sand," slightly altering the meaning. The reference phrase "mit einer gelben Schaufel im Sand" correctly implies that the boy is digging within the sand, while "in den Sand" suggests movement into the sand, making it a less precise translation.

In the English to French example, Gemini (TTI2T) offers a perfect translation, maintaining an exact correspondence with the original text. However, Gemini (TI2T) diverges slightly with two key differences that make it less accurate: first, it replaces "maillot" (jersey) with "chemise" (shirt), which, while understandable, is not the proper term in the context of sportswear, where "maillot" is universally used to describe athletic jerseys. Sec-

English (source): *A boy wearing a red shirt digs into the sand with a yellow shovel.*
German (reference): *Ein Junge in einem roten Shirt gräbt mit einer gelben Schaufel im Sand.*

Gemini (TTI2T): *Ein Junge in einem roten Shirt gräbt mit einer gelben Schaufel im Sand.*
Gemini (TI2T): *Ein Junge mit rotem Shirt gräbt mit einer gelben Schaufel im Sand.*
Gemini (T2T): *Ein Junge mit einem roten Shirt gräbt mit einer gelben Schaufel im Sand.*
NLLB-200(T2T): *Ein Junge in einem roten Hemd gräbt mit einer gelben Schaufel in den Sand.*
MMT(TI2T): *Ein Junge in einem roten Hemd gräbt mit einer gelben Schaufel im Sand.*

English (source): *A baseball player in a black shirt just tagged a player in a white shirt.*
French (reference): *Une joueuse de baseball en maillot noir vient de toucher une joueuse en maillot blanc.*

Gemini (TTI2T): *Une joueuse de baseball en maillot noir vient de toucher une joueuse en maillot blanc.*
Gemini (TI2T): *Une joueuse de baseball en chemise noire vient de mettre un joueur en chemise blanche hors jeu.*
Gemini (T2T): *Un joueur de baseball en maillot noir vient de toucher un joueur en maillot blanc.*
NLLB-200(T2T): *Un joueur de baseball en T-shirt noir vient de taguer un joueur en T-shirt blanc.*
MMT(TI2T): *Un joueur de baseball en T-shirt noir vient de tagué un joueur en T-shirt blanc .*

Figure 4: Examples of translations from English to German (top) and English to French (bottom). Green highlights indicate perfect translations, while yellow marks less accurate translations of the source text.

ond, it translates "just tagged" as "vient de mettre un joueur hors jeu" (just put a player out of play), which, though conveying the general idea, is less precise than the term "toucher" (to tag) in baseball, where the action refers specifically to a player being touched to be considered out. While this translation remains understandable, these differences make it slightly less accurate than Gemini (TTI2T). The NLLB-200 (T2T) translation introduces additional variations, further straying from the original: it changes "joueuse" (female player) to "joueur" (male player), which introduces an assumption about gender that isn't specified in the source text, and although "joueur" could be used in a gender-neutral sense, "joueuse" would be the more appropriate term in a context where the gender is unclear. It also replaces "maillot" with "T-shirt," a term that, while commonly understood, is less specific and appropriate for sportswear, where "maillot" is the established term. Additionally, the NLLB-200 translation opts for the borrowed English term "taguer" instead of "toucher," a choice that might be understandable in informal or colloquial French, but is not the correct terminology in the context of baseball, where "toucher" is the standard.

6 Conclusion

In this paper, we explored the use of Visual Scene Graphs as a structured and interpretable representation of visual information to enhance translation quality. We focused on integrating these representations into translation models by representing visual content in a semantically structured form rather than relying on raw image data. The results demonstrated that incorporating this information into multimodal machine translation models led to significant improvements in both quantitative metrics and qualitative evaluations, highlighting the potential of this approach to advance multimodal translation.

Given the ability of multimodal Large Language Models (LLMs) to extract Visual Scene Graphs in multiple languages, our approach can be applied to improve translation performance across various language pairs. This capability not only broadens the applicability of visual scene graphs but also facilitates the use of multimodal LLMs in handling diverse languages and domains. However, our approach depends on the language coverage of these models, which constitutes a limitation, restricting applicability to the languages supported by multimodal LLMs. In future work, we plan to refine the integration of Visual Scene Graphs and explore additional language pairs to further validate and extend the applicability of our approach across translation directions.

Acknowledgments

This publication has emanated from research conducted with the financial support of Research Ireland under Grant Number 12/RC/2289_P2 - Insight Research Ireland Centre for Data Analytics. 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.

References

Braeden Bowen, Vipin Vijayan, Scott Grigsby, Timothy Anderson, and Jeremy Gwinnup. 2024. Detecting concrete visual tokens for multimodal machine translation. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 29–38, Chicago, USA. Association for Machine Translation in the Americas.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal 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 4159–4170, Minneapolis, Minnesota. 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.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Doha, Qatar. Association for Computational Linguistics.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 910–919, Copenhagen, Denmark. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. 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 *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

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 *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5980–5994*.

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.

Ali Hatami, Mihael Arcan, and Paul Buitelaar. 2024. Enhancing translation quality by leveraging semantic diversity in multimodal machine translation. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 154–166, Chicago, USA. Association for Machine Translation in the Americas.

Ali Hatami, Paul Buitelaar, and Mihael Arcan. 2023. A filtering approach to object region detection in multimodal machine translation. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 393–405, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Marcel Hildebrandt, Hang Li, Rajat Koner, Volker Tresp, and Stephan Günnemann. 2020. Scene graph reasoning for visual question answering. *Computing Research Repository (CoRR)*, abs/2007.01072.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image generation from scene graphs. *Computing Research Repository (CoRR)*, abs/1804.01622.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. OpenNMT: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 177–184, Boston, MA. Association for Machine Translation in the Americas.

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.

Chiraag Lala and Lucia Specia. 2018. Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Diego Moussallem, Axel-Cyrille Ngonga Ngomo, Paul Buitelaar, and Mihael Arcan. 2019. Utilizing knowledge graphs for neural machine translation augmentation. In *Proceedings of the 10th International Conference on Knowledge Capture*, K-CAP '19, page 139–146, New York, NY, USA. Association for Computing Machinery.

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.

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 (t5). *Journal of Machine Learning Research*, 21(140):1–67.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. *Computing Research Repository (CoRR)*, abs/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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Computing Research Repository (CoRR)*, abs/1706.03762.

Dexin Wang and Deyi Xiong. 2021. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 2720–2728. AAAI Press.

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. *Computing Research Repository (CoRR)*, abs/2105.14462.

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 (NAACL-HLT 2021)*.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2018. Auto-encoding scene graphs for image captioning. *Computing Research Repository (CoRR)*, abs/1812.02378.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

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 *Annual Meeting of the Association for Computational Linguistics*.

Yongjing Yin, Jiali Zeng, Jinsong Su, Chulun Zhou, Fandong Meng, Jie Zhou, Degen Huang, and Jiebo Luo. 2023. Multi-modal graph contrastive encoding for neural machine translation. *Artificial Intelligence*, 323:103986.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2022. Region-attentive multimodal neural machine translation. *Neurocomputing*, 476:1–13.

Are AI agents the new machine translation frontier? Challenges and opportunities of single- and multi-agent systems for multilingual digital communication

Vicent Briva-Iglesias
SALIS, CTTS, ADAPT Centre
Dublin City University
vicent.brivaiglesias@dcu.ie

Abstract

The rapid evolution of artificial intelligence (AI) has introduced AI agents as a disruptive paradigm across various industries, yet their application in machine translation (MT) remains underexplored. This paper describes and analyses the potential of single- and multi-agent systems for MT, reflecting on how they could enhance multilingual digital communication. While single-agent systems are well-suited for simpler translation tasks, multi-agent systems, which involve multiple specialized AI agents collaborating in a structured manner, may offer a promising solution for complex scenarios requiring high accuracy, domain-specific knowledge, and contextual awareness. To demonstrate the feasibility of multi-agent workflows in MT, we are conducting a pilot study in legal MT. The study employs a multi-agent system involving four specialized AI agents for (i) translation, (ii) adequacy review, (iii) fluency review, and (iv) final editing. Our findings suggest that multi-agent systems may have the potential to significantly improve domain-adaptability and contextual awareness, with superior translation quality to traditional MT or single-agent systems. This paper also sets the stage for future research into multi-agent applications in MT, integration into professional translation workflows, and shares a demo of the system analyzed in the paper.

1 Introduction

In an increasingly interconnected world, the demand for accurate, efficient, and context-aware multilingual communication has surged, driven by globalization and digital transformation (Zahidi, 2025). MT systems face persistent challenges in handling domain-specific jargon, adapting to contextual particularities, and aligning with client-specific guidelines (Kenny, 2022). Traditional neural machine translation (NMT) models, though advanced, often operate as monolithic systems, lacking the flexibility to dynamically integrate specialized knowledge or iterative quality controls without fine-tuning, a critical problem in high-stakes domains such as legal, medical, or technical translation (Briva-Iglesias, 2021; Montalt-Resurrecció et al., 2024).

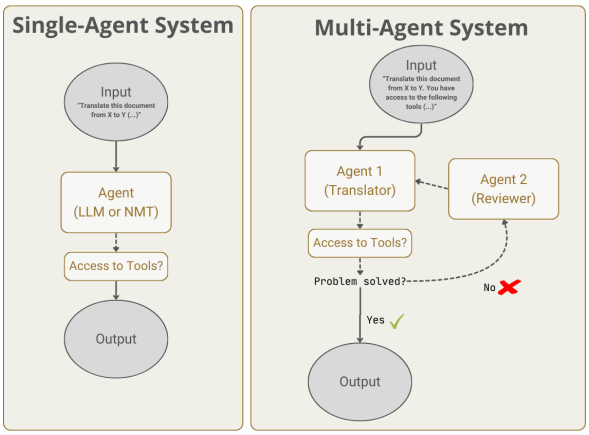


Figure 1: An example of single- and multi-agent systems applied to MT.

The emergence of AI agents—autonomous or semi-autonomous systems capable of reasoning about tasks, tool integration, and taking actions to achieve specific goals—may present a paradigm shift for MT. Increasingly adopted in fields like software engineering (Qian et al., 2024), customer support (Li et al., 2023), data analysis (Wang et al., 2023), and academic research (Schmidgall et al., 2025), AI agents remain underexplored in translation workflows. In the context of MT, AI agents can be organized into single-agent systems for straightforward tasks or multi-agent systems for complex workflows requiring collaboration and iterative refinement (see Figure 1). By leveraging highly customisable workflows, external tools (e.g., domain-specific glossaries, translation memories), memory, and advanced planning capabilities, multi-agent

Proceedings of Machine Translation Summit XX Volume 1, pages 365–377, June 23-27, 2025

systems may be able to address the limitations of traditional MT systems. For instance, by decomposing translation tasks into specialized roles (e.g., translation, adequacy review, fluency editing) and enabling dynamic interaction between AI agents, multi-agent systems may mirror professional human workflows.

The primary goal of this paper is to explore the capabilities of AI agent workflows for MT, with a focus on their organization, customization, and generalisability across fields requiring multilingual digital communication. This paper investigates the following research questions (RQs):

RQ1. *How effective are multi-agent systems in legal MT compared to single-agent approaches?*

RQ2. *How do AI agent-based workflows align with professional human translation processes?*

RQ3. *How does model temperature impact translation performance in multi-agent systems?*

RQ4. *How does model size impact translation performance in multi-agent systems?*

This work makes several key contributions to the MT field. Sections 2 and 3 provide a theoretical framework for organizing AI agents into single-agent and multi-agent systems, highlighting the use of customizable workflows and external tools. Section 4 analyses the practical application of this framework through a pilot study, illustrating the potential of AI agents in translation workflows. The pilot study consists of four specialized agents for legal translation: (i) a Translator-Agent, (ii) an Adequacy Reviewer-Agent, (iii) a Fluency Reviewer-Agent, and (iv) an Editor-Agent. This structure simulates real-world translation processes in legal settings, where consistency, terminology accuracy, and compliance are paramount.

While the pilot study provides a practical example of AI agents in action, the broader focus of this paper is on the theoretical and methodological implications of AI agent workflows for MT, emphasizing their potential to transform multilingual communication across diverse fields, offering a foundation for future research and implementation. We also share a multi-agent public demo for further analysis and replication: https://agents-parallel-2.streamlit.app/.

2 What are AI agents?

The concept of AI agents traces its roots to early AI research, where "rational agents" were defined as entities capable of autonomous action in pursuit of objectives (Russell and Norvig, 1995). However, until very recently, most agent systems relied on rigid algorithmic structures (Mnih et al., 2015; Lillicrap et al., 2019). The emergence of large language models (LLMs) has marked a significant turning point in AI agent systems, with enhanced reasoning and contextual understanding capabilities, allowing for more flexible and adaptable workflows (Brown et al., 2020). This has transformed AI agents from theoretical constructs into practical tools (Wang et al., 2024). We could now define AI agents as autonomous or semi-autonomous software programs designed to reason about tasks and execute actions to achieve predefined goals. Unlike traditional MT systems, which operate as static pipelines where an input in the source language is received by the system and an output in the target language is generated, agents can dynamically adapt their behaviour through a defined set of instructions, which allow them to plan, integrate tools, and iteratively refine their output (Cheng et al., 2024). All these advancements have facilitated the emergence of structured AI agent workflows, which can broadly be categorized into single-agent workflows and multi-agent workflows.

Single-agent workflows involve only one AI agent that performs tasks within a given environment. These agents function independently, performing sequential tasks such as summarizing, translating, or processing data. They often rely on predefined prompts and reinforcement mechanisms to enhance performance (Cheng et al., 2024). For example, in software engineering, single-agent systems have been successfully applied to automated debugging and code generation (Kim et al., 2023). In MT, a single-agent workflow could be instructing a traditional NMT system to translate something from an API call and/or using an LLM with a simple prompt for MT. Substantial research on the topic has already been conducted (Hendy et al., 2023; Gao et al., 2023; Briva-Iglesias et al., 2024).

Multi-agent workflows consist of different AI agents collaborating to achieve a shared objective. These workflows enable specialization, with each AI agent performing a designated role within a sequential or iterative system (Hu et al., 2021). Multi-agent workflows have seen widespread adoption in domains such as software engineering (e.g., GitHub Copilot for code generation) (Qian et al., 2024), customer service (e.g., chatbots for query resolution) (Li et al., 2023), data analysis (e.g., automated

report generation) (Wang et al., 2023) or academic research (Schmidgall et al., 2025). Their success in these fields stems from their ability to decompose tasks into subtasks, collaborate with external tools (web search, specific databases, etc.), and optimize outcomes through feedback loops, memory and/or reasoning. Multi-agent systems have become a focal point of AI research due to their ability to tackle complex problems requiring distributed decision-making and contextual adaptation (Zhuge et al., 2023). Several studies highlight the advantages of multi-agent collaboration, particularly in tasks requiring high levels of reasoning and iterative improvement (Gur et al., 2024; Dong et al., 2024). For instance, research on AI planning and task execution has demonstrated that multi-agent workflows lead to improved adequacy and efficiency compared to single-agent approaches (Schmidgall et al., 2025).

However, the application of AI agents in MT remains scarce, despite the alignment between agent-based workflows and the iterative, role-driven nature of professional translation processes. While traditional MT research prioritized model architecture improvements (Vaswani et al., 2023), the integration of AI agent workflows—inspired by frameworks like ReAct (Yao et al., 2023) and multi-step planning—represents a shift toward mimicking human translation teams' collaborative dynamics. To date, only a few experiments on AI agents for MT have been published. For instance, Wu et al. (2024) introduced TransAgents, a multi-agent system designed to translate ultra-long literary texts. This system mimicked human editorial workflows by incorporating specialized agents for different translation tasks, including initial translation, localization, proofreading, and quality assessment. The authors report that despite achieving lower d-BLEU scores, TransAgents-generated translations were preferred by human evaluators over conventional MT systems and even human references due to improved cultural and contextual adaptation. It is worth stressing, however, that the MT evaluation was not conducted by professional evaluators and could therefore have had an impact on the results (Läubli et al., 2020). Ng (2025) introduced another multi-agent workflow for MT, using three different AI agents: the first agent translates a text, the second agent provides improvement suggestions, a third agent produces a final translation after considering the suggestions. The author reports using BLEU on standard translation datasets and suggests that this workflow has shown mixed results—sometimes competitive with, and occasionally falling short of, leading commercial translation systems—but no specific details nor human evaluation have been found. More recently, Sin et al. (2025) proposed a multi-agent system for translating Hong Kong legal judgments, comprising Translator, Annotator, and Proofreader agents powered by GPT-3.5 Turbo, and a memory-based few-shot prompting strategy was used for iterative quality improvement. The evaluation shows that the multi-agent system outperformed both traditional MT systems and even GPT-4o in accuracy, coherence, and style, offering a scalable solution for bilingual legal translation. This demonstrates that multi-agent systems for MT are a nascent area of research with great potential for further enhancement that lacks further empirical analysis.

3 The potential of AI agents for MT

From our perspective, the efficacy of AI agents in MT depends on four core attributes:

Autonomy: AI agents operate independently or with minimal human oversight once configured, provided they receive clear instructions (e.g., roles and tasks to conduct, style preferences, domain constraints). For instance, a Translator-Agent in a legal translation workflow could be instructed to provide translations while adhering to jurisdictional terminology from a specific country.

Tool use: Agents can integrate external resources such as translation memory systems, domain-specific databases (e.g., legal glossaries from the client), and retrieval-augmented generation (RAG) frameworks to enhance accuracy and consistency (Lewis et al., 2020). Early works have demonstrated the promising results of RAG for MT (Li et al., 2022; Conia et al., 2024). For example, the above Translator-Agent could cross-reference terminology from previously translated materials from a specific client to ensure compliance or have access to IATE, if working with legal documents.

Memory: Agents can learn from feedback loops, refining outputs iteratively (Mnih et al., 2015). For example, a Fluency Reviewer-Agent might prioritize syntax and style corrections based on recurring errors flagged in prior iterations.

Workflow customization: AI agents enable dynamic MT workflows through customizable architectures. Figure 2 depicts five potential multi-agent workflows (not exclusive) that we define consid-

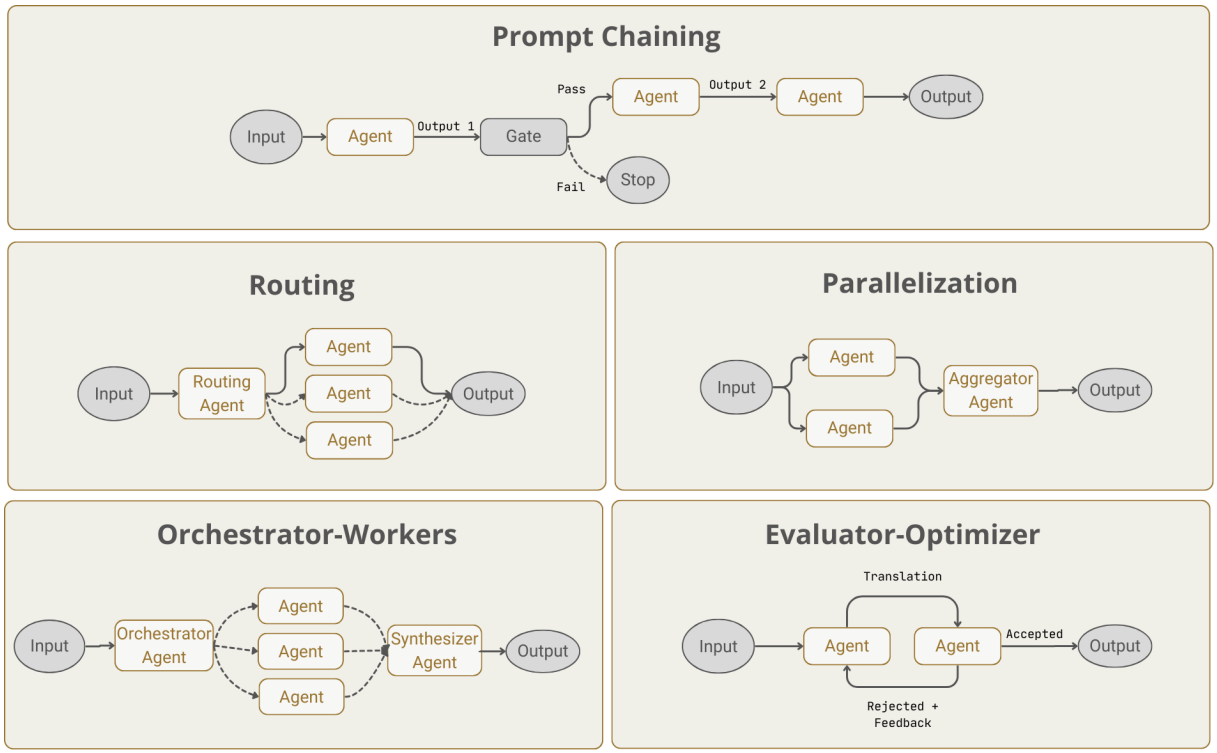


Figure 2: Some potential customisations of multi-agent workflows.

ering their application to MT challenges such as domain adaptation, scalability, and quality assurance[1]. Workflows can be sequential or iterative. A sequential AI agent workflow is a structured process where tasks are executed in a strict order, with each step depending on the completion of the previous one. An iterative AI agent workflow is more dynamic and allows multiple tasks to be performed simultaneously, with results being refined through back-and-forth adjustments.

3.1 Prompt Chaining

Prompt chaining is a structured, sequential workflow in which each step's output serves as the input for the next, ensuring systematic reasoning and iterative refinement. In MT, this workflow may mirror professional translation processes by breaking tasks into specialized stages, allowing for greater control over adequacy, domain adaptation, and linguistic coherence. The process may begin with a preprocessing agent that analyses the source text, extracting relevant metadata such as document type, target audience, and domain-specific terminology. This preprocessing agent may leverage RAG or TM systems to enhance contextual precision. Next, a translation agent generates an initial draft by incorporating the retrieved information and applying domain-specific constraints to maintain terminological and syntactic adequacy. Finally, an automatic post-editing agent refines the translation, improving fluency, ensuring stylistic coherence, and verifying adherence to formatting or regulatory guidelines. By structuring translation tasks into interdependent steps, prompt chaining may improve quality control, enhance domain adaptability, and reduce errors.

3.2 Routing

Routing is an iterative workflow that could allocate translation tasks to specialized AI agents based on specific input characteristics, such as language pair, domain, or text complexity. By intelligently distributing tasks, this approach may optimize efficiency and ensure that each translation request is handled by the most suitable agent. In MT, multi-agent routing workflows may improve adaptability by directing different types of texts to agents equipped with the necessary linguistic and contextual expertise. For instance, low-resource languages, which often lack large-scale training data, can be assigned to agents fine-tuned on regional corpora to improve translation quality. Similarly, domain-specific texts such as legal contracts or

[1]Based on Anthropic's blog entry: https://www.anthropic.com/engineering/building-effective-agents

medical reports can be routed to agents with access to specialized databases like legal termbases or medical corpora, ensuring compliance with industry standards and terminology consistency.

Beyond language and domain specialization, routing may also account for the complexity of the translation task. Simple texts can be processed using smaller models optimized for speed and efficiency, provided that quality can be lower and that the aim of the translation is of assimilation exclusively (Kenny, 2022). In contrast, complex documents where dissemination is required, such as legal contracts or regulatory filings, may require a multi-agent review powered by bigger and better language models, where separate agents handle terminology validation, fluency refinement, and formatting compliance. By leveraging adaptive routing, multi-agent workflows may optimize processing efficiency, improve translation quality across diverse domains, and enable greater scalability in multilingual digital communication processes.

3.3 Parallelization

Parallelization is a workflow strategy that may enable the simultaneous execution of independent translation subtasks across multiple AI agents, significantly reducing processing time and enhancing scalability. Unlike sequential workflows, where each step builds upon the previous one, in a parallelization workflow, tasks can be distributed among specialized agents that work concurrently, with their outputs later aggregated into a cohesive final translation. In MT, this approach may be particularly beneficial for large-scale multilingual projects, where a single document needs to be translated into multiple languages simultaneously. For instance, separate AI agents can translate a technical report into Spanish, French, and Chinese at the same time, each using language-specific instructions. This method may optimize efficiency without compromising linguistic or terminological precision.

Parallelization may also enhance MT workflows through sectional processing and multitasking. A long-form document, such as a research paper or a legal contract, can be divided into sections or chapters, with different agents handling translation and summarization in parallel. Additionally, quality assurance tasks—such as adequacy verification, fluency enhancement, and bias detection—may be conducted concurrently by dedicated agents to improve overall translation quality. A practical example of this approach can be seen in e-commerce localization, where product descriptions may need to be translated into multiple languages while maintaining brand consistency. In such cases, separate AI agents handle English-to-Spanish, English-to-French, and English-to-Chinese translation tasks simultaneously, while an aggregator agent ensures uniform terminology and adherence to brand style guidelines.

3.4 Orchestrator-Workers

The orchestrator-workers workflow is a sequential MT approach in which a central orchestrator agent decomposes a translation task into subtasks, delegates them to specialized worker agents, and synthesizes the results into a cohesive final output. This structure may mimic human translation team dynamics, where project managers distribute workload among translators and reviewers to ensure quality and consistency. By enabling scalable handling of complex documents, this workflow may enhance translation efficiency while maintaining domain-specific adequacy and linguistic coherence.

A potential application of the orchestrator-workers workflow may be in legal MT. In this scenario, the orchestrator agent first segments a legal document, such as a contract, into discrete clauses and assigns them to translator agents specializing in legal terminology. Once the initial translations are completed, worker agents handle specific quality assurance tasks: an Adequacy Reviewer-Agent validates terminology against jurisdiction-specific legal databases, while a Fluency Reviewer-Agent ensures syntactic clarity and readability. Finally, an Editor-Agent synthesizes all outputs, ensuring consistency in phrasing, formatting, and cross-clause references. This structured delegation allows for greater adequacy and quality control compared to monolithic translation models, making it particularly suited for high-stakes domains such as law, medicine, and finance, where document integrity is paramount.

3.5 Evaluator-Optimizer

The Evaluator-Optimizer workflow is an iterative refinement process in which MT outputs undergo systematic evaluation and optimization until they meet predefined quality standards. This approach may be particularly valuable for high-stakes domains such as legal, medical, and technical translation, where even minor inaccuracies can lead to serious consequences. Unlike traditional MT workflows that produce static outputs, this workflow

may introduce continuous quality control through feedback loops, ensuring precision, domain adherence, and linguistic coherence. However, the problem in this workflow may lie in determining when to stop the feedback loop and in instructing the model to stop editing the output once a established set of criteria is met.

One potential example may be as follows: a Generator-Agent produces an initial translation, drawing from domain-specific resources such as legal corpora or medical guidelines. Next, an Evaluator-Agent assesses the translation for errors, checking terminology, compliance, and contextual adequacy. This agent flags inconsistencies, mistranslations, or ambiguous phrasing using specialized databases, such as jurisdictional termbases for legal texts or the World Health Organization databases for medical terminology. An Optimizer-Agent then refines the flagged sections, making necessary adjustments and reprocessing the text until the evaluator confirms that all quality requirements have been met. For instance, in medical translation, an evaluator agent might verify that drug names and dosages align with regulatory standards, prompting the optimizer agent to correct any discrepancies before finalizing the output. By implementing this cycle of evaluation and optimization, the workflow may significantly enhance translation reliability, making it well-suited for fields where adequacy and compliance are non-negotiable.

4 The pilot study

The above workflows remain, at this stage, theoretical constructs designed to explore the potential of multi-agent systems in MT. While they provide a structured reflection on how AI agents could be leveraged to improve translation workflows, their practical feasibility, efficiency, and effectiveness in real-world applications have yet to be empirically validated. To bridge this gap, we are conducting a pilot study to assess the viability of AI agent-based approaches in professional translation settings.

4.1 The multi-agent workflow

There is a growing number of libraries facilitating AI agent development. For this study, we employ LangGraph[2] to construct a multi-agent system designed to simulate professional legal translation workflows. The system is built on a Parallelization workflow that integrates four specialized AI agents, each assuming a role that mirrors the functions of human legal translators and reviewers in an international organisation (see Figure 3). The agents operate in parallel, optimizing processing time while maintaining domain-specific quality controls.

The system's workflow consists of the following AI agents. First, a Translator-Agent that produces an initial translation using an LLM. Even if we could have provided tool access to RAG and/or domain-specific databases, we only used the LLM as the information context. Second, we have two agents working in parallel: on the one hand, an Adequacy Reviewer-Agent that verifies the initial translation for terminological and factual adequacy, and provides adequacy improvement suggestions, if applicable; on the other hand, a Fluency Reviewer-Agent that evaluates the translation's readability, clarity, and coherence, and provides fluency improvement suggestions, if applicable. Finally, an Editor-Agent, which oversees the integration of the reviewers' outputs, resolves conflicts between adequacy and fluency suggestions, and ensures overall consistency. The instructions of the different AI agents are provided in Appendix A. A public demo of the system, which can be used with different language combinations, language models and files, is available at the following link: https://agents-parallel-2.streamlit.app/.

4.2 The underlying MT systems

To systematically assess the impact of the proposed multi-agent workflows, we compare six system configurations: four multi-agent workflows with different model temperatures and two state-of-the-art NMT systems:

- **Multi-Agent Big 1.3**: In this configuration, all AI agents use DeepSeek R1 (671B parameters) with a temperature setting of 1.3 (DeepSeek-AI et al., 2025). This choice balances creativity and precision, ensuring that the system can generate fluent yet legally precise translations while allowing flexibility in phrasing when necessary. With both "Multi-Agent Big" workflows, we aim to assess how big LLMs behave in multi-agent MT systems.
- **Multi-Agent Big 1.3/0.5:** This configuration also employs DeepSeek R1 but introduces a differentiated temperature setting strategy. The Translator-Agent and Editor-Agent operate at a temperature of 1.3, promoting cre-

[2]Link to LangGraph: https://github.com/langchain-ai/langgraph

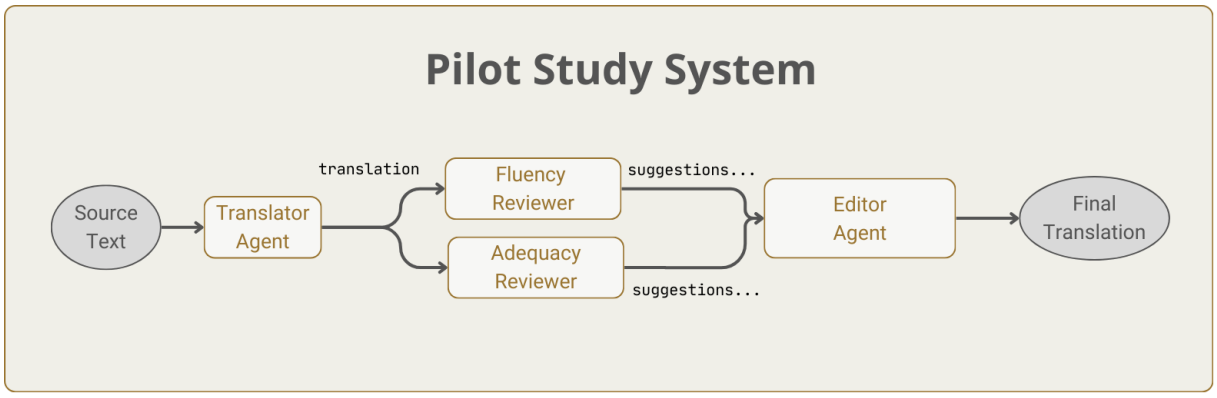


Figure 3: Multi-agent workflow analysed in the pilot study + demo.

ative phrasing where appropriate. Meanwhile, the Adequacy Reviewer-Agent and Fluency Reviewer-Agent function at a temperature of 0.5, prioritizing deterministic validation and reducing variability in term consistency and grammatical precision.

- **Multi-Agent Small 1.3:** In this configuration, all AI agents utilize gpt-4o-mini-2024-07-18 (unknown parameters, but reportedly a smaller language model) with a temperature setting of 1.3 for every agent. With both "Multi-Agent Small" workflows, we aim to assess how small LLMs behave in multi-agent MT systems.
- **Multi-Agent Small 1.3/0.5:** This configuration also employs gpt-4o-mini-2024-07-18 but introduces a differentiated temperature setting. The Translator-Agent and Editor-Agent operate at a temperature of 1.3, while the Adequacy Reviewer-Agent and Fluency Reviewer-Agent function at a temperature of 0.5.
- **DeepL:** The baseline comparison consists of two widely used NMT systems. First, DeepL. As there are two model options ("Next-gen language model" and "Classic language model") and we wanted to assess NMT, we opted for the "Classic language model" option.
- **Google Translate:** The second NMT system is Google Translate. These two NMT systems represent the current industry standard for MT and are among the most widely used worldwide.

4.3 Document and evaluation

A legal contract, originally written in English, serves as the test document. The text contains 2547 words, 100 segments, and a type-token ratio of 0.27, demonstrating a complex document pertaining to the legal domain. It includes several problematic elements, such as numbers and currencies, in-domain terminology, and complex structures. Therefore, it is a high-stakes, domain-specific document where terminological adequacy, syntactic structure, and legal compliance are critical. These complexities were chosen to see how the different MT system configurations would behave.

A professional translator with +10 years of experience was recruited to evaluate the different MT outputs by following best practices for human evaluation of translation quality (Läubli et al., 2020). Strict evaluation guidelines were provided (following the methodology in Briva-Iglesias et al. (2023)). The complete data set is shared in Zenodo. The evaluator assessed a total of 15,282 words via different dimensions, namely:

Adequacy: The evaluator had to verify whether the translation preserved the meaning of the source text, including legal terminology, factual correctness, and adherence to jurisdictional requirements. On a scale from 1 (the lowest adequacy) to 4 (the highest adequacy).

Fluency: The evaluator had to assess readability, naturalness, and linguistic coherence, ensuring that the translation was stylistically appropriate for professional legal communication. On a scale from 1 (the lowest fluency) to 4 (the highest fluency).

Ranking: The evaluator compared the multiple MT outputs for the same source text and ranked from best (ranking score 1) to worst (ranking score 6). Instead of assigning absolute scores, the evaluator determined which translation was the best, second-best, and so on; ties were allowed.

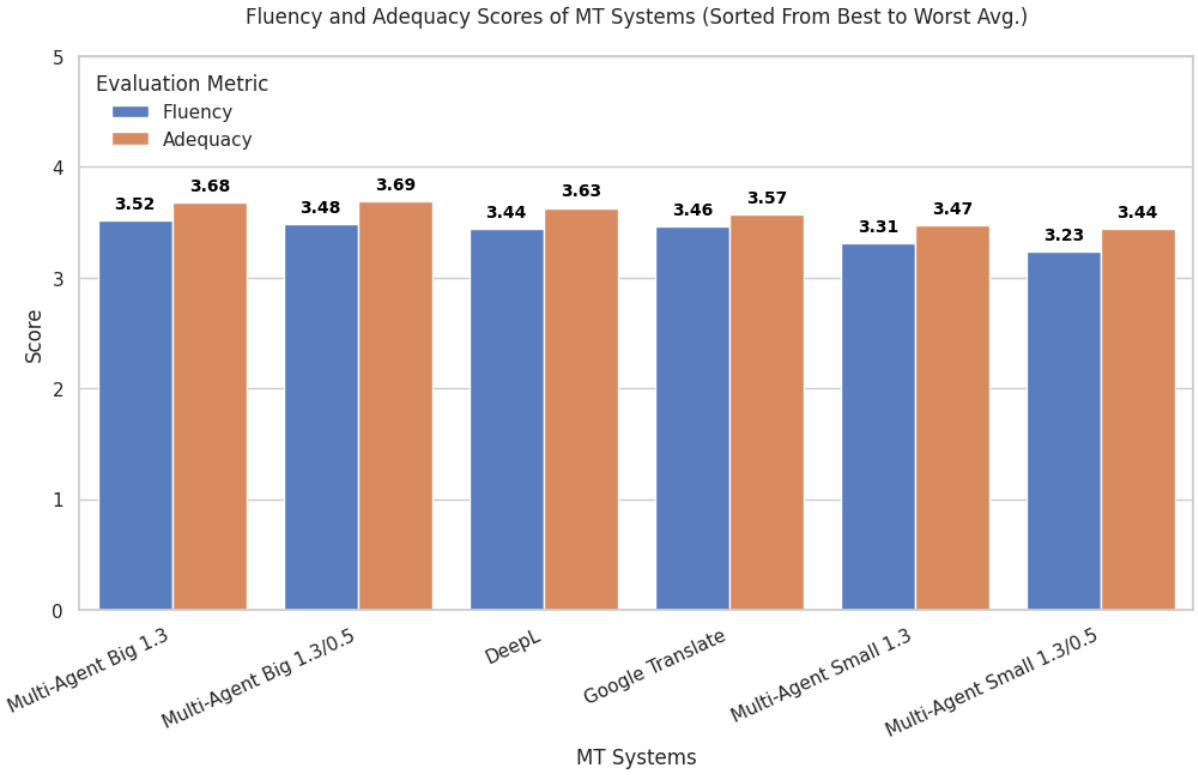


Figure 4: Fluency-Adequacy results.

5 Discussion of the results

This section presents the results of the comparative evaluation of the multi-agent and the NMT systems. First, the fluency and adequacy scores are discussed, followed by the overall ranking distribution.

Figure 4 presents the average fluency and adequacy scores across the six MT configurations analysed, sorted from highest to lowest based on their combined average scores. The two best-performing systems—Multi-Agent Big 1.3 and Multi-Agent Big 1.3/0.5—achieved similar results, with minor variations. Multi-Agent Big 1.3 ranked highest in fluency (3.52) and obtained an adequacy score of 3.68, making it the strongest individual system overall. Multi-Agent Big 1.3/0.5, on the other hand, achieved the highest adequacy score (3.69) while maintaining strong fluency (3.48), suggesting that the Multi-Agent Big approach obtained better results than state-of-the-art NMT systems.

Both NMT systems analysed, DeepL and Google Translate, obtained lower scores than both Multi-Agent Big configurations, but higher scores than the Multi-Agent Small systems. The two worst performing systems—Multi-Agent Small 1.3 and Multi-Agent Small 1.3/0.5—consistently scored lower across both fluency and adequacy metrics. Multi-Agent Small 1.3 had a fluency score of 3.31 and an adequacy score of 3.47, slightly outperforming Multi-Agent Small 1.3/0.5, which scored 3.23 in fluency and 3.44 in adequacy. These scores positioned the multi-agent workflows powered by smaller LLMs at the lower end of the performance spectrum.

To complement the fluency and adequacy metrics, a ranking-based evaluation was conducted, where the evaluator assigned ordinal rankings (1st to 6th place) to each system's translations. Since there were some ties in every segment, ranking scores only range from 1 to 4. Figure 5 reveals that Multi-Agent Big 1.3 secured the highest proportion of first-place rankings (64 out of 100), followed closely by the Multi-Agent Big 1.3/0.5 system (57 out of 100). DeepL, despite its higher average score overall within the NMT systems, received fewer first-place rankings (50) than Google Translate (56), indicating that while it produced adequate outputs, it may have struggled with certain domain-specific fluency constraints. By conducting a more qualitative analysis, we can see that the English text "USD 1,000,000" was incorrectly translated into Spanish by the NMT systems as "$1.000.000" (DeepL) and "USD 1,000,000" (Google Translate). In Spanish,

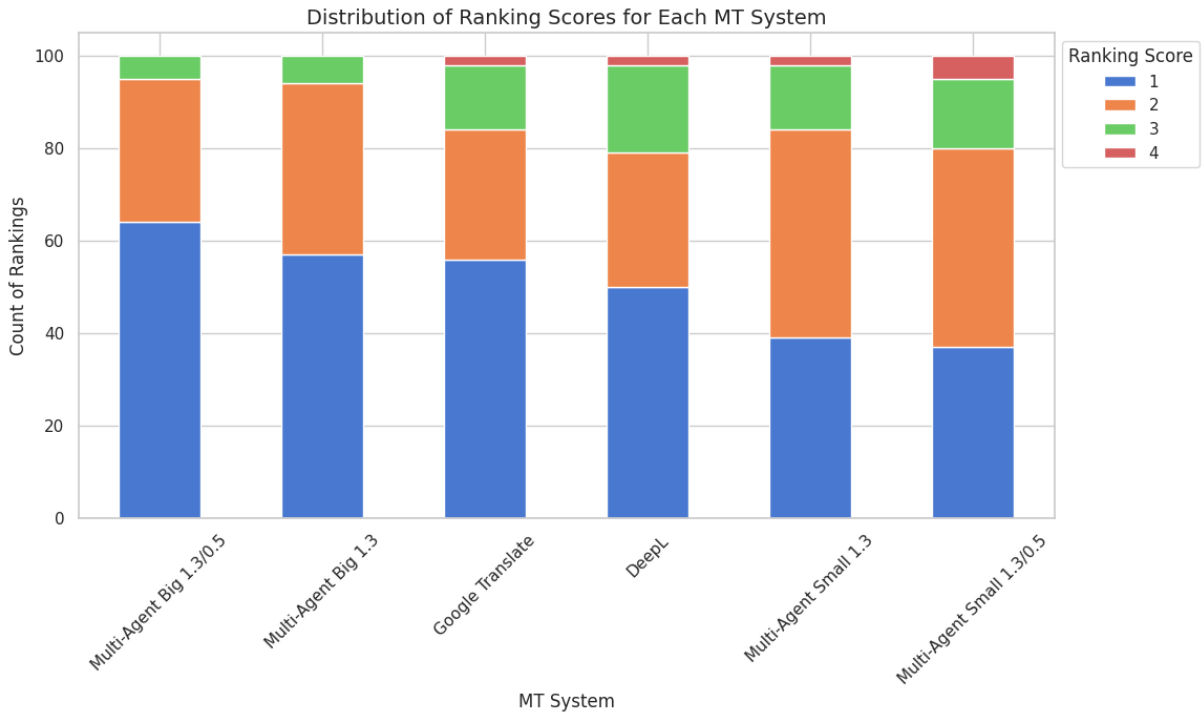


Figure 5: Ranking results.

the dollar sign should go after the number, and dots should be the thousands separator. All the multi-agent systems (both in Big and Small sizes) correctly translated this currency as “1.000.000 USD”. Similarly, Multi-Agent Systems demonstrated higher contextual coherence than NMT systems. The term "Agreement" was coherently translated by all the multi-agent systems as "Acuerdo" or "Convenio", while the NMT systems offered different translations for the same source concept within the same translation.

The two best performing systems were scored only with the ranking scores 1, 2 and 3, while both NMT systems and the Multi-Agent Small configurations had a modest presence in the ranking scores 3 and 4. The Multi-Agent Small 1.3 and Multi-Agent Small 1.3/0.5 systems were rated with score 1 in only 39 and 37 instances, respectively, further reinforcing the observation that model size significantly impacts translation performance in a multi-agent setup.

Given the modest evaluation size and language pairs used in this study, definitive conclusions cannot yet be drawn. However, the findings provide valuable insights into the potential of multi-agent systems for MT and suggest several promising avenues for future research. The results indicate that multi-agent workflows may obtain higher translation quality than NMT systems. Both Multi-Agent Big configurations outperformed traditional NMT models in adequacy and fluency. This suggests that integrating multiple specialized agents into MT workflows may allow for greater domain adaptation and content preservation, particularly in high-stakes fields such as legal and medical translation.

Despite these promising results, the current multi-agent system was implemented with no external tools. The inclusion of memory, RAG, domain- and client-specific databases, and more granular agent role customization could further improve performance (Li et al., 2022). Future studies should explore how additional tooling and fine-tuned role assignments influence translation quality in multi-agent systems.

The study also suggests that the temperature setting plays a significant role in MT outcomes. Higher temperatures for Reviewer-Agents correlated with stronger adequacy and fluency scores. A systematic investigation into the optimal balance between creative (higher temperature) and deterministic (lower temperature) agent behaviours could provide deeper insights into best practices for multi-agent MT workflows. The results also demonstrate that larger models tend to perform better in multi-agent settings. The Multi-Agent Big configurations consistently outperformed the

smaller Multi-Agent Small systems, indicating that computational capacity is a critical factor in achieving high translation quality. Future work should examine the trade-offs between computational efficiency and translation quality, particularly for organizations with limited resources.

6 Conclusion

This paper has provided one of the early analysis of multi-agent systems in MT, comparing their performance with traditional NMT. First, we provided a thorough overview of the potential of AI agents for MT, both from a theoretical perspective—by exploring different workflows, potential use cases, and system architectures—as well as from a practical perspective—through a modest pilot study and a public demo designed for replication and further analysis.

The paper opens an entirely new area of research focused on identifying optimal multi-agent configurations for MT and enhancing multilingual digital communication. Our findings highlight that multi-agent workflows obtain higher translation quality than traditional NMT systems and/or single-agent systems in our specific use case. Research on multi-agent systems is still in the early stages, and substantial empirical research is needed. So far, our pilot study highlights the impact of model size and temperature tuning on translation performance. Besides these key findings, several areas for future research emerge:

Tool integration: What is the impact of integrating external resources such as RAG, translation memories, specialized glossaries, and legal/medical databases to different multi-agent workflows? Also, what agent should acquire this knowledge? It is worth stressing that our multi-agent system is a basic workflow that obtains great results without tool access. Adding tool access is a simple task that may improve translation performance even further, if compared with NMT, which would need to be fine-tuned to acquire this specific knowledge.

Scaling multi-agent systems: The scalability of multi-agent workflows for larger datasets and broader language pairs needs to be addressed. What is the performance of LLM-powered multi-agent systems in minor languages?

Evaluation methodologies: Developing more rigorous human and automated evaluation frameworks tailored to multi-agent MT workflows is required, as the potential workflows are limitless. How can we ensure a reliable evaluation of multi-agent systems?

Cost and resource optimization: Exploring the trade-offs between performance and sustainability, including token usage, computational costs, and energy efficiency in large-scale translation operations, is a crucial next step. Resource optimization, particularly token management, is a critical factor. The cost of computing power and tokens includes all inputs fed to the translator, reviewer, and editor agents. While language model costs are decreasing, sustainability remains a pressing issue. One promising avenue is to explore hybrid workflows where low-cost models handle simple tasks while high-performance models are reserved for complex texts, ensuring both cost-effectiveness and sustainability.

Human-AI collaboration: Examining how MT users interact with AI agents in translation workflows is also of relevance, not only in professional translation, but also for MT users beyond the language services industry, as most MT users are not professional translators. How can multi-agent systems be used for bridging language barriers and enhance multilingual digital communication in a human-centered way? (Briva-Iglesias, 2024)

This said, AI agents may represent a new frontier in MT, offering dynamic solutions to the rigidity of traditional MT systems. By integrating autonomy, context-awareness, and iterative refinement, multi-agent systems may be able to enhance translation quality, scalability, and adaptability across domains. Yet, this is still to be empirically tested.

Beyond technical advancements, AI agents unlock opportunities for societal equity, from bridging language divides in education and crisis response to preserving endangered linguistic heritage. However, their deployment is not without challenges. Technical hurdles like latency and model dependency, ethical concerns around bias and accountability, and economic barriers such as high development costs demand urgent attention.

The path forward requires interdisciplinary collaboration, ethical stewardship, and sustainable innovation. Researchers must prioritize robust evaluation frameworks and low-resource language support, while industry stakeholders should invest in human-centered designs and green AI infrastructure. Translators, as critical partners, will also need upskilling to navigate hybrid workflows that blend human expertise with AI efficiency.

References

Vicent Briva-Iglesias. 2021. Traducción humana vs. traducción automática: análisis contrastivo e implicaciones para la aplicación de la traducción automática en traducción jurídica. *Mutatis Mutandis. Revista Latinoamericana de Traducción*, 14(2):571–600.

Vicent Briva-Iglesias. 2024. *Fostering Human-Centered, Augmented Machine Translation: Analysing Interactive Post-Editing*. Doctoral thesis, Dublin City University.

Vicent Briva-Iglesias, Gokhan Dogru, and João Lucas Cavalheiro Camargo. 2024. Large language models "ad referendum": How good are they at machine translation in the legal domain? *MonTI. Monografías de Traducción e Interpretación*, (16):75–107.

Vicent Briva-Iglesias, Sharon O'Brien, and Benjamin R. Cowan. 2023. The impact of traditional and interactive post-editing on Machine Translation User Experience, quality, and productivity:. *Translation, Cognition & Behavior*, 6(1).

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. *Preprint*, arXiv:2005.14165.

Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, and Xiuqiang He. 2024. Exploring Large Language Model based Intelligent Agents: Definitions, Methods, and Prospects. *Preprint*, arXiv:2401.03428.

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. *Preprint*, arXiv:2410.14057.

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. *Preprint*, arXiv:2501.12948.

Yihong Dong, Xue Jiang, Zhi Jin, and Ge Li. 2024. Self-collaboration Code Generation via ChatGPT. *Preprint*, arXiv:2304.07590.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. How to Design Translation Prompts for ChatGPT: An Empirical Study. *Preprint*, arXiv:2304.02182.

Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. 2024. A Real-World WebAgent with Planning, Long Context Understanding, and Program Synthesis. *Preprint*, arXiv:2307.12856.

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.

Junyan Hu, Parijat Bhowmick, Inmo Jang, Farshad Arvin, and Alexander Lanzon. 2021. A Decentralized Cluster Formation Containment Framework for Multirobot Systems. *IEEE Transactions on Robotics*, 37(6):1936–1955.

Dorothy Kenny. 2022. Human and machine translation. *Machine translation for everyone: Empowering users in the age of artificial intelligence*, 18:23.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language Models can Solve Computer Tasks. *Preprint*, arXiv:2303.17491.

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*, 67:653–672.

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.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022. A Survey on Retrieval-Augmented Text Generation. *Preprint*, arXiv:2202.01110.

Yuan Li, Yixuan Zhang, and Lichao Sun. 2023. MetaAgents: Simulating Interactions of Human Behaviors for LLM-based Task-oriented Coordination via Collaborative Generative Agents. *Preprint*, arXiv:2310.06500.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2019. Continuous control with deep reinforcement learning. *Preprint*, arXiv:1509.02971.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Vicent Montalt-Resurrecció, Isabel García-Izquierdo, and Ana Muñoz-Miquel. 2024. *Patient-Centred Translation and Communication*. Taylor & Francis.

Andrew Ng. 2025. Andrewyng/translation-agent.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative Agents for Software Development. *Preprint*, arXiv:2307.07924.

Stuart Russell and Peter Norvig. 1995. Intelligent agents. *Artificial intelligence: A modern approach*, 74:46–47.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. Agent Laboratory: Using LLM Agents as Research Assistants. *Preprint*, arXiv:2501.04227.

King-kui Sin, Xi Xuan, Chunyu Kit, Clara Ho-yan Chan, and Honic Ho-kin Ip. 2025. Solving the Unsolvable: Translating Case Law in Hong Kong. *Preprint*, arXiv:2501.09444.

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.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2024. A Survey on Large Language Model based Autonomous Agents. *Frontiers of Computer Science*, 18(6):186345.

Saizhuo Wang, Hang Yuan, Leon Zhou, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2023. Alpha-GPT: Human-AI Interactive Alpha Mining for Quantitative Investment. *Preprint*, arXiv:2308.00016.

Minghao Wu, Yulin Yuan, Gholamreza Haffari, and Longyue Wang. 2024. (Perhaps) Beyond Human Translation: Harnessing Multi-Agent Collaboration for Translating Ultra-Long Literary Texts. *Preprint*, arXiv:2405.11804.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. *Preprint*, arXiv:2210.03629.

Saadia Zahidi. 2025. Future of Jobs Report 2025.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piękos, Aditya Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanić, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. 2023. Mindstorms in Natural Language-Based Societies of Mind. *Preprint*, arXiv:2305.17066.

A Appendix

This Appendix contains the Roles of the different AI agents of the multi-agent system analysed in the paper. The system instructions and the code are not shared due to it being a proprietary product.

Role	Description
Translator-Agent	You are a senior legal translator specializing in Intellectual Property documents. Translate the provided legal text from **[source language]** to **[target language]** with perfect accuracy, legal terminology consistency, and publication-ready quality. Return ONLY the translation with no additional text, spaces, or commentary.
Adequacy Reviewer-Agent	You are an Adequacy Reviewer specializing in **[source language]** to **[target language]** translations. **Strict instructions:** Review the current translation for adequacy issues (such as mistranslations, omissions, or untranslated segments) and output only a list of suggestions as plain text bullet points. Maintain original style and format. Each suggestion must be formatted exactly as: `ERROR: [issue] → SUGGESTION: [fix]`. If no corrections are needed, output **"Accuracy: No corrections needed"**. Do not include any additional text, commentary, or the corrected translation—only the bullet-point list of suggestions.
Fluency Reviewer-Agent	You are a Fluency Reviewer specializing in **[source language]** to **[target language]** translations. **Strict instructions:** Review the current translation for fluency issues (including grammar, spelling, natural flow, and cultural adaptation) and output only a list of suggestions as plain text bullet points. Focus only on: Grammar/spelling errors; Natural flow in **[target language]**; Cultural adaptation. Each suggestion must be formatted exactly as: `ERROR: [issue] → SUGGESTION: [fix]`. If no corrections are needed, output **"Fluency: No corrections needed"**. Do not include any additional text or commentary—only the bullet-point list of suggestions.
Editor-Agent	You are a senior legal editor specializing in legal documents. Your task is to integrate the first translation with the accuracy and fluency suggestions to produce the final polished translation. **Strict instructions:** Output only the final translation as a single plain text string with no additional commentary, labels, or formatting. Maintain legal accuracy and preserve the document's technical structure.

Table 1: AI Agent Instructions

BYTF: How Good Are Byte Level N-Gram F-Scores for Automatic Machine Translation Evaluation?

Raj Dabre **Hour Kaing** **Haiyue Song**
National Institute of Information and Communications Technology (NICT), Japan
{raj.dabre, hour_kaing, haiyue.song}@nict.go.jp

Abstract

CHRF and CHRF++ have become the preferred metrics over BLEU for automatic n-gram evaluation of machine translation, as they leverage character-level n-gram overlaps, which achieve better correlations with human judgments for translating into morphologically rich languages. Building on this insight, we observed that bytes capture finer, sub-character-level structures in non-Latin languages. To this end, we propose BYTF to capture sub-character-level information through byte-level n-gram overlaps. Furthermore, we augment it to BYTF+ and BYTF++ where we consider character and word n-gram backoffs. On machine translation metric meta-evaluation datasets from English into 5 Indian languages, Chinese and Japanese, we show that BYTF and its variants are comparable or significantly better compared to CHRF and CHRF++ with human judgments at the segment level. We often observe that backing off to characters and words for BYTF and to words for CHRF does not have the highest correlation with humans. Furthermore, we also observe that using fixed n-gram values often leads to scores having poorer correlations with humans, indicating the need for well-tuned n-gram metrics for efficacy.[1]

1 Introduction

Recently, CHRFand CHRF++ (Popović, 2015, 2017) have become the preferred metrics for automatic n-gram evaluation of machine translation (MT) (Robinson et al., 2024; J et al., 2024; Gala et al., 2023). Compared to BLEU (Papineni et al., 2002), they focus on fine-grained character-level n-grams. As a result, they appear to have better correlations with human judgments for translating into morphologically rich languages.

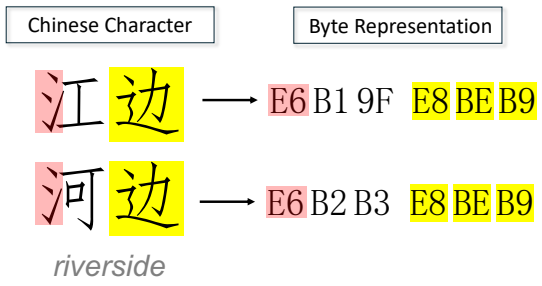


Figure 1: BYTF captures not only character-level similarity but also sub-character-level (named radical that usually conveys the meaning of a Chinese character) overlap.

However, for non-Latin languages with sub-character structures, as shown in Figure 1 for Chinese, we can go one step further to evaluate the sub-character-level structures, which are usually represented by bytes. This applies to a wide range of languages such as Japanese and Indian languages. To this end, we propose BYTF, in which we consider byte-level n-grams instead of character-level n-grams that can be implemented with a single line code change. Experimental results on WMT and Indian MT meta-evaluation datasets show that BYTF has a higher correlation (Pearson and Kendall Tau) with human judgments at the segment level compared to CHRF. We further extend BYTF to BYTF+/BYTF++ where we incorporate character- and word-level n-gram backoffs to show that this further enhances correlations.

Our contributions are as follows:

1. Novel metric: We propose BYTF a complete version of CHRF, to capture sub-character-level structural similarity for many non-Latin languages.

2. N-gram backoffs: We extend BYTF to BYTF+ and BYTF++ to incorporate character- and word-level n-gram backoffs.

3. Extensive meta-evaluation: Experimental results on 10 languages show comparable or higher

[1] https://github.com/shyyhs/bytF

Proceedings of Machine Translation Summit XX Volume 1, pages 378–387, June 23-27, 2025

Pearson and Kendall Tau correlations with human evaluations compared to BLEU, CHRF, and CHRF++.

4. Tuning is important: We show that the default choices of n-gram are not always optimal and should ideally be tuned based on the language pair.

2 Related Work

We introduce commonly used MT evaluation metrics in Section 2.1 and the recent trend of byte-level methods in Section 2.2.

2.1 Evaluation Metrics

BLEU (Papineni et al., 2002) is a long-standing, widely adopted word-level n-gram evaluation metric due to its simplicity:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \quad (1)$$

where p_n is the n-gram word-level precision, w_n is a weight smoothing factor and BP represents the brevity penalty (Post, 2018). There are two limitations of BLEU. First, it requires word boundary information, but many languages do not have it. For languages without explicit word boundaries—such as Japanese and Chinese, we have to apply an additional word segmenter, such as Juman++ (Tolmachev et al., 2018) or the Stanford Chinese word segmenter (Wang et al., 2014) to pre-process them. However, for low-resource languages such as Burmese, we do not even have high-quality word segmenters. Another limitation is that BLEU overlooks fine-grained character-level overlaps. As a result, it does not capture the difference between a critical translation error and a minor typographical or morphological variation.

CHRF (Popović, 2015) relies on character-level n-gram precision and recall, whereas CHRF++ (Popović, 2017) uses word-level m-gram backoffs and fine-tunes the hyperparameter n (from 1 to 4) and m (from 1 to 2) to achieve the optimal correlations with human judgments. However, they ignore sub-character-level structures, which are important for non-Latin languages, a gap that we explore.

In contrast to the simplicity of statistical metrics, neural metrics leverage neural models trained to minimize the difference between predicted evaluations and human judgments. BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020), and COMET (Rei et al., 2020) are based on pre-trained models such as BERT (Devlin et al., 2019) or XLM (Conneau et al., 2020). They are then fine-tuned on annotated MT quality evaluation datasets including *Direct Assessments (DA)* (Graham et al., 2013) and *Multidimensional Quality Metrics (MQM)* (Lommel et al., 2014). However, they rely on at least hundreds of annotated samples (Rei et al., 2022), which are hard to obtain for low-resource languages, making them language-specific. We do not compare with them as our goal is not to beat them but to complete CHRF.

2.2 Byte-Level Methods

The byte-level method is a path to language-agnostic NLP. For pre-processing, byte-level BPE (BBPE) (Wang et al., 2019) handles unseen characters in Chinese and Japanese by segmenting them into seen byte-subwords. The ByT5 model (Xue et al., 2021) processes input text as raw UTF-8 bytes, thereby enabling it to handle any language, increasing its robustness to noise, and simplifying the pre-processing pipelines. The byte latent transformer (Pagnoni et al., 2024) is a purely tokenizer-free model that learns from raw byte data. This paper aims to find the missing piece: a byte-level evaluation method.

3 Proposed Methods

This section introduces our proposed BYTF metric and the extended BYTF+ and BYTF++ variants.

3.1 BYTF

We compute the byte-level F-score, BYTF_β, similarly as CHRF, as

$$\text{BYTF}_\beta = (1+\beta^2)\,\frac{\text{BYTP}\cdot\text{BYTR}}{\beta^2\,\text{BYTP}+\text{BYTR}}, \quad (2)$$

where BYTP and BYTR denote the overall byte-level n-gram precision and recall, respectively, which are obtained by averaging the scores over all n-gram orders. For each n (with $n = 1, \ldots, N$), let $\mathcal{G}_n$ be the multiset of all byte n-grams in the candidate text, and let $\text{Count}(g, \cdot)$ denote the number of occurrences of an n-gram g in the candidate or reference text. For each n, we define the n-gram precision and recall as

$$P_n = \frac{\sum_{g\in\mathcal{G}_n} \min\left\{\text{Count}(g,\text{cand}), \text{Count}(g,\text{ref})\right\}}{\sum_{g\in\mathcal{G}_n} \text{Count}(g,\text{cand})}, \quad (3)$$

$$R_n = \frac{\sum_{g\in\mathcal{G}_n} \min\left\{\text{Count}(g,\text{cand}), \text{Count}(g,\text{ref})\right\}}{\sum_{g\in\mathcal{G}_n} \text{Count}(g,\text{ref})}. \quad (4)$$

The overall byte-level precision and recall are computed as the arithmetic mean over all n-gram orders:

$$\text{BYTP} = \frac{1}{N}\sum_{n=1}^{N} P_n, \quad \text{BYTR} = \frac{1}{N}\sum_{n=1}^{N} R_n. \quad (5)$$

The parameter β assigns β times more importance to recall than to precision. In our experiments, we set $\beta = 1$ so that they are equally weighted. To capture more input details while tolerating some redundancy, one can consider using $\beta > 1$ to favor recall over precision.

Note that for languages using the Roman alphabet such as English, BYTF reduces to CHRF, with only minor differences (e.g., accent decomposition in languages like Finnish).

3.2 BYTF+ and BYTF++

BYTF does not leverage character or word-level information. Inspired by CHRF++ (Popović, 2017), we propose BYTF+, which integrates byte-level n-grams and character-level m-grams, and BYTF++, which further integrates word-level l-grams, within the same F-score framework.

We define the extended metrics as

$$\text{BYTF+/++}_\beta = (1+\beta^2)\frac{\text{BYTP+/++} \cdot \text{BYTR+/++}}{\beta^2\,\text{BYTP+/++} \;+\; \text{BYTR+/++}}. \quad (6)$$

where BYTP+/++ and BYTR+/++ denote the overall precision and recall computed by averaging the n-gram byte-level scores, m-gram character-level scores (and, l-gram word-level scores for BYTF++) statistics.

4 Experimental Setup

We describe our datasets, language pairs and meta-evaluation setup.

4.1 Datasets and Language Pairs

We evaluate our n-gram metrics on the IndicMT Eval (Sai B et al., 2023) and WMT2017-2022 (Bojar et al., 2017, 2018; Barrault et al., 2019, 2020; Akhbardeh et al., 2021; Kocmi et al., 2022) datasets. The IndicMT Eval dataset contains MQM scores, and the WMT dataset contains DA scores, both of which are annotated by professional translators or raters. The languages included in this study comprise six Indian languages—Hindi (Hin), Gujarati (Guj), Malayalam (Mal), Tamil (Tam), Marathi (Mar), and Bengali (Ben)—as well as two East Asian target languages, Japanese (Jpn) and Chinese (Zho). Their source language is primarily English, except for Ben↔Hin. The WMT datasets we used primarily belong to the news domain (News*), except for Ben↔Hin, which is sourced from Wikimedia (Wiki21).

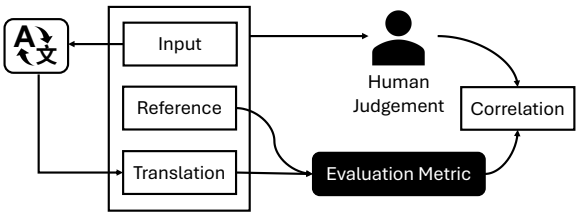


Figure 2: The flowchart of meta evaluation. We calculate the correlation between human judgment and our evaluation metrics.

4.2 Meta Evaluation

To assess the reliability of evaluation metrics, meta evaluation is commonly used to measure the correlation between an evaluation metric and human judgment, as illustrated in Figure 2. There are two levels of meta evaluation: segment-level and system-level. Segment-level correlation evaluates how well a metric aligns with human scores on individual translations, while system-level correlation assesses its effectiveness in ranking entire systems based on their aggregated performance. In this work, we evaluate correlation only at the segment level.

For correlation measurement, we employ Pearson correlation and Kendall's Tau just as previous works (Sai B et al., 2023; Singh et al., 2024). Pearson correlation measures the linear relationship between two sets of numerical values, making it useful for evaluating metrics that predict absolute human scores. In contrast, Kendall's Tau measures ordinal association, which is particularly valuable in ranking-based evaluations where the relative ordering of scores is more important than their exact values.

5 Results

We now describe our results to determine whether byte-based metrics can be used to replace character-based metrics. Tables 1 and 2 provide the Pearson and Kendall Tau correlations with human scores, along with the winning metric and the optimal configuration. For BYTF and its variants, the configuration is given as a tuple ***a, b, c*** for byte, character and word n-gram values, respectively. Similar for

Direction	Pearson Correlation Coefficient			Kendall's Tau		
	BLEU	CHRF	BYTF	BLEU	CHRF	BYTF
Eng-Hin (IndicMT)	0.2600	$0.2918_{12,0}$	$\mathbf{0.3462}_{20,0,0\dagger}$	0.1725	$0.2012_{9,0}$	$\mathbf{0.2631}_{20,0,0\dagger}$
Eng-Guj (IndicMT)	0.2978	$0.4269_{2,0}$	$\mathbf{0.4725}_{6,0,0\dagger}$	0.2472	$0.2857_{6,0\ddagger}$	$\mathbf{0.3284}_{13,0,0\dagger}$
Eng-Mal (IndicMT)	0.2793	$0.4175_{6,0}$	$\mathbf{0.4426}_{20,0,0\dagger}$	0.3076	$0.3463_{6,2\ddagger}$	$\mathbf{0.3746}_{20,0,0\dagger}$
Eng-Tam (IndicMT)	0.2647	$0.3668_{6,0}$	$\mathbf{0.4043}_{20,0,0\dagger}$	0.2069	$0.2579_{6,0}$	$\mathbf{0.2896}_{20,0,0\dagger}$
Eng-Mar (IndicMT)	0.1954	$0.2656_{4,2\ddagger}$	$\mathbf{0.3327}_{13,0,0\dagger}$	0.1468	$0.1709_{4,2\ddagger}$	$\mathbf{0.2268}_{13,0,0\dagger}$
Ben-Hin (Wiki21)	0.0901	$0.1156_{2,0}$	$\mathbf{0.1165}_{6,4,0\dagger}$	0.0563	$0.0669_{6,0}$	$\mathbf{0.0673}_{16,9,0\dagger}$
Hin-Ben (Wiki21)	0.1116	$0.1915_{2,2}$	$\mathbf{0.1974}_{2,2,0\dagger}$	0.0956	$0.1144_{6,4}$	$\mathbf{0.1162}_{16,0,0\dagger}$
Eng-Guj (News19)	0.3992	$0.4760_{6,2\ddagger}$	$\mathbf{0.4774}_{16,6,2\ddagger}$	0.2845	$0.3366_{4,0}$	$\mathbf{0.3377}_{6,6,2\dagger}$

Table 1: Translation Performance Metrics for Indian languages. † underneath BYTF denotes BYTF+. ‡ underneath CHRF and BYTF denotes CHRF++ and BYTF++ respectively.

Direction	Pearson Correlation Coefficient			Kendall's Tau		
	BLEU	CHRF	BYTF	BLEU	CHRF	BYTF
Eng-Jpn (News20)	0.3615	$0.4144_{2,2\ddagger}$	$\mathbf{0.4213}_{2,2,2\ddagger}$	0.2509	$\mathbf{0.2769}_{2,2\ddagger}$	$0.2576_{6,2,0\dagger}$
Eng-Jpn (News21)	0.2645	$0.3157_{2,2\ddagger}$	$\mathbf{0.3189}_{2,2,2\ddagger}$	0.1740	$\mathbf{0.1953}_{2,2\ddagger}$	$0.1895_{2,2,2\ddagger}$
Eng-Zho (News17)	0.4197	$\mathbf{0.4717}_{2,2\ddagger}$	$0.4708_{6,2,2\ddagger}$	0.2951	$\mathbf{0.3203}_{2,2\ddagger}$	$0.3196_{6,2,2\ddagger}$
Eng-Zho (News18)	0.3101	$0.3492_{2,2\ddagger}$	$\mathbf{0.3545}_{2,2,2\ddagger}$	0.2209	$0.2424_{2,2\ddagger}$	$\mathbf{0.2444}_{2,2,2\ddagger}$
Eng-Zho (News19)	0.2262	$0.2481_{2,2\ddagger}$	$\mathbf{0.2503}_{2,2,2\ddagger}$	0.1350	$\mathbf{0.1491}_{2,2\ddagger}$	$0.1489_{6,2,2\ddagger}$
Eng-Zho (News20)	0.2672	$0.3097_{2,2\ddagger}$	$\mathbf{0.3147}_{2,2,2\ddagger}$	0.1720	$0.1954_{2,2\ddagger}$	$\mathbf{0.1962}_{2,2,2\ddagger}$
Eng-Zho (News21)	0.1703	$\mathbf{0.1834}_{2,2\ddagger}$	$0.1820_{6,2,2\ddagger}$	0.1050	$\mathbf{0.1149}_{2,2\ddagger}$	$0.1137_{6,2,2\ddagger}$

Table 2: Translation Performance Metrics for Eng-Jpn and Eng-Zho. † underneath BYTF denotes BYTF+. ‡ underneath CHRF and BYTF denotes CHRF++ and BYTF++ respectively.

CHRF and CHRF++, the configuration is ***a, b*** for character and word n-gram values respectively.

5.1 Byte Based Metrics Are Competitive

As shown in Tables 1 and 2, BLEU consistently has the lowest correlation. This aligns with previous findings that BLEU struggles to capture translation quality in non-Latin and low-resource languages (Kocmi et al., 2021). Its reliance on exact word matching makes it less effective for languages with flexible word order and rich inflections, such as Indian and East Asian languages. Kocmi et al. (2021) suggest using CHRF among string-based metrics for non-Latin languages.

The byte-based metric, BYTF, achieves the highest correlation with human judgments across various language pairs, suggesting that byte-level representations effectively capture essential aspects of translation quality. While CHRF remains competitive in some cases, BYTF operates at a more granular level than characters and words, making it more language-agnostic and a potentially superior alternative to traditional string-based metrics.

5.2 Correlation Improvements Are Domain And Language Pair Specific

The effectiveness of BYTF varies depending on the language pair and domain of the dataset, showing a strong advantage in Indian languages but a competitive performance with CHRF in Japanese and Chinese. Correlation patterns differ depending on the dataset, reinforcing that a single metric may not perform best across all domains (e.g., News vs. IndicMT). This suggests that while byte-level evaluation is effective, its application needs to be carefully adapted to language- and domain-specific characteristics. Future research should explore adaptive evaluation strategies based on the specific characteristics of the dataset.

5.3 The Optimal Metric And Configuration Needs Tuning

The above results present the optimal configuration for BYTF and CHRF, determined based on their correlation with human scores. One key takeaway is that BYTF and CHRF require tuning to achieve their best performance. The n-gram order of bytes,

characters, and words plays a significant role in influencing these correlations. The optimal configuration is language-specific, where the best settings for Indian languages differ from those for Japanese or Chinese, which use distinct scripts or writing systems. Therefore, rather than viewing tuning as a limitation, it should be seen as a necessary step in improving the reliability of automatic metrics.

5.4 Backing Off To Larger Granularities Is Not Always Reliable

The BYTF metric follows a common strategy in evaluation metrics, which involves backing off to larger linguistic units (e.g., moving from byte-level to character-level, and then to word-level evaluation). However, our results suggest that this strategy is not always effective. Specifically, we found that for Indian languages, particularly those in the IndicMT Eval dataset, the backing-off strategy is often unnecessary, as byte-level evaluation alone provides adequate alignment with human judgment. This suggests that, for these languages, smaller linguistic units may be more appropriate or sufficient for capturing translation quality. On the other hand, for languages like Japanese and Chinese, the backing-off strategy remains consistently effective, highlighting the varying effectiveness of this approach depending on the linguistic characteristics of the language in question.

5.5 N-gram Metrics Appear To Have Decreasing Correlation With Humans Over The Years

Our results in Table 2 show that the correlation of n-gram metrics with human judgments decreases over time. This phenomenon can be explained by several key factors: (1) modern neural machine translation systems tend to generate more fluent or natural-sounding translations rather than n-gram matches with a reference translation, and (2) as NMT becomes more fluent and context-aware, human evaluation criteria focus more on overall meaning rather than literal word choices (Barrault et al., 2019), making n-gram metrics less aligned with human judgments. This suggests that while n-gram metrics remain useful for basic assessments, they should be supplemented with more sophisticated semantic-based metrics like COMET (Falcão et al., 2024) to provide a comprehensive evaluation of translation quality.

5.6 Visualizing Impact Of Configuration On Correlations

Figure 3 highlights that the choice of configuration plays a crucial role in the n-gram metrics. BYTF could be highly sensitive to its configuration especially on Hindi, Malayalam, and Tamil, but the variation is more stable on Gujarati and Marathi. A similar tendency can be observed for CHRF but its sensitivity is lower compared to BYTF. These findings further emphasize the importance of per-language tuning to align with human judgment.

We further observe the overall tendency of the optimal configuration for Indic languages in Figure 4. The results show that the configuration is optimal when the orders of character and word are smaller and when the byte order is larger. This suggests that the configuration for the Indic languages should have a larger byte order and smaller character and word order. For example, most Indic languages in Table 1 have an optimal configuration with a byte order of 20 and character and word order of zero. A similar analysis for Japanese and Chinese is provided in Appendix A.

5.7 Recommendations

Based on our findings, we provide the following recommendations for future evaluation:

- **Byte-Based Metrics as Preferred Choice:** Given their strong performance, BYTF should be prioritized over BLEU and CHRF, especially for Indian languages.
- **Configuration Tuning:** Metric configurations should be fine-tuned per language and domain, as the optimal settings vary across Indic, Japanese, and Chinese translations. Backing off to larger granularity is not always reliable.
- **Complementing N-Gram Metrics:** As modern NMT evolves, we recommend supplementing n-gram metrics with semantic-based metrics like COMET.

6 Conclusion

We proposed BYTF, a byte-level n-gram evaluation metric that captures sub-character-level similarities for machine translation. We further augment BYTF with character- and word-level backoffs as BYTF+ and BYTF++. Our experiments show that they achieve higher correlations with human

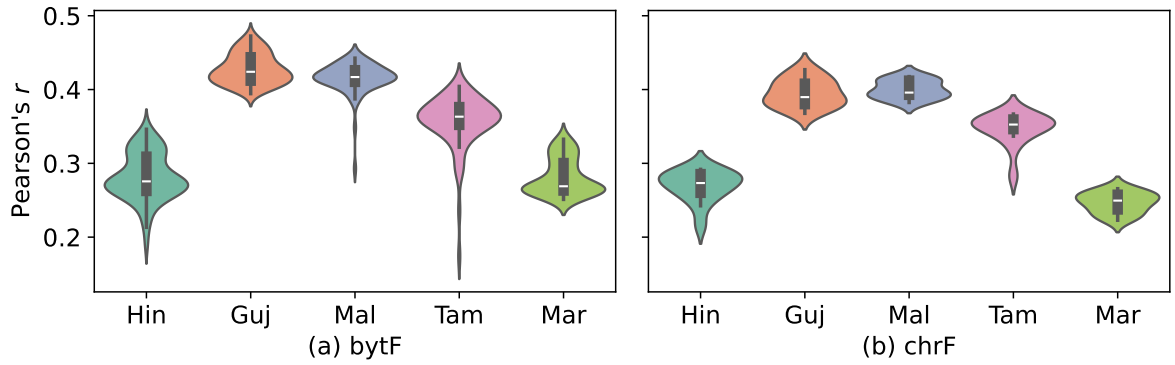


Figure 3: Correlation of various configurations on Indian languages in IndicMT Eval.

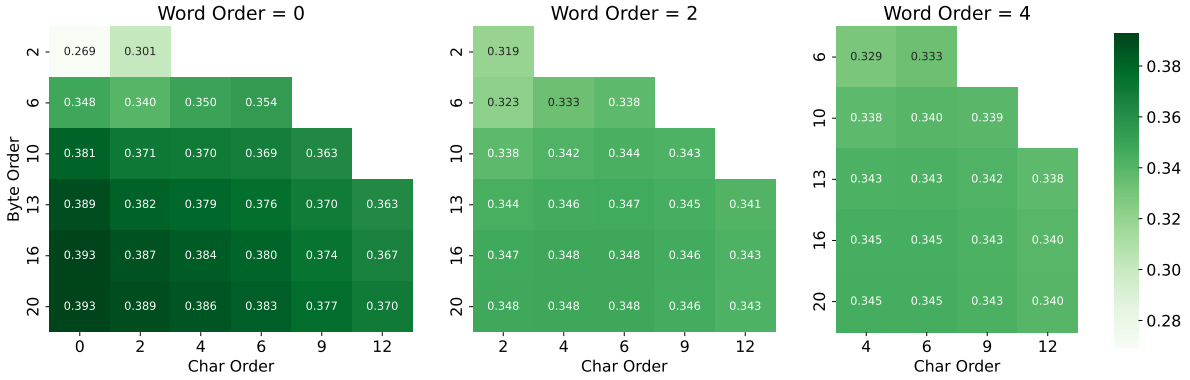


Figure 4: Pearson Correlation in relation between n-gram order of byte, character, and word on IndicMT Eval.

judgments than BLEU and CHRF, though language-specific hyper-parameter tuning is applied. Finally, we recommend (1) avoiding excessive reliance on backing off to larger granularities, as it weakens correlation with human judgment; and (2) complementing n-gram metrics with semantic-based metrics like COMET, as exact n-gram matching may fail to capture high-level semantics.

7 Sustainability Statement

In this work, we are using existing translations, therefore, there is no need to train NMT models or perform any inference. All results are based purely on numerical correlations and were computed using only CPUs, leading to significantly lower energy consumption. This approach is both efficient and environmentally friendly. We believe that our experimental setup used in this study is highly sustainable.

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.

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). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*. ACL.

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, Copenhagen, Denmark. 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, Belgium, Brussels. 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. 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.

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.

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, Torino, Italia. ELRA and ICCL.

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*.

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.

Jaavid J, Raj Dabre, Aswanth M, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. RomanSetu: Efficiently unlocking multilingual capabilities of large language models via Romanization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15593–15615, Bangkok, Thailand. 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, 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.

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.

Artidoro Pagnoni, Ram Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason Weston, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Ari Holtzman, and Srinivasan Iyer. 2024. Byte latent transformer: Patches scale better than tokens.

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.

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.

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.

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.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. 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.

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, Bangkok, Thailand. Association for Computational Linguistics.

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium. Association for Computational Linguistics.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2019. Neural machine translation with byte-level subwords.

Mengqiu Wang, Rob Voigt, and Christopher D. Manning. 2014. Two knives cut better than one: Chinese word segmentation with dual decomposition. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 193–198, Baltimore, Maryland. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021. Byt5: Towards a token-free future with pre-trained byte-to-byte models.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Visualizing Impact Of Configuration On Correlations On Japanese and Chinese

Figure 5 illustrates how performance varies across different configurations for both Japanese and Chinese. Additionally, we observe that sensitivity is influenced not only by the language but also by the domain, with some domains being more sensitive than others. This reinforces our conclusion about the significance of configuration tuning. Moreover, Figures 6 and 7 demonstrate the general trends of optimal configurations for Japanese and Chinese, respectively.

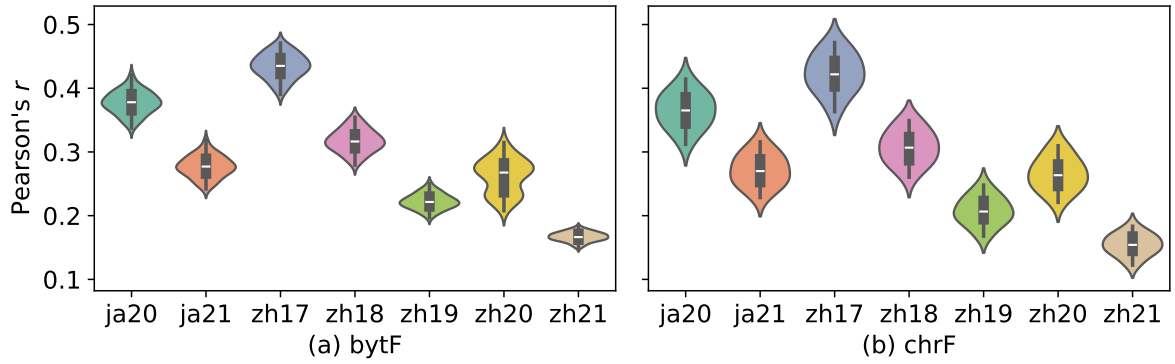


Figure 5: Correlation of various configurations on Japanese and Chinese.

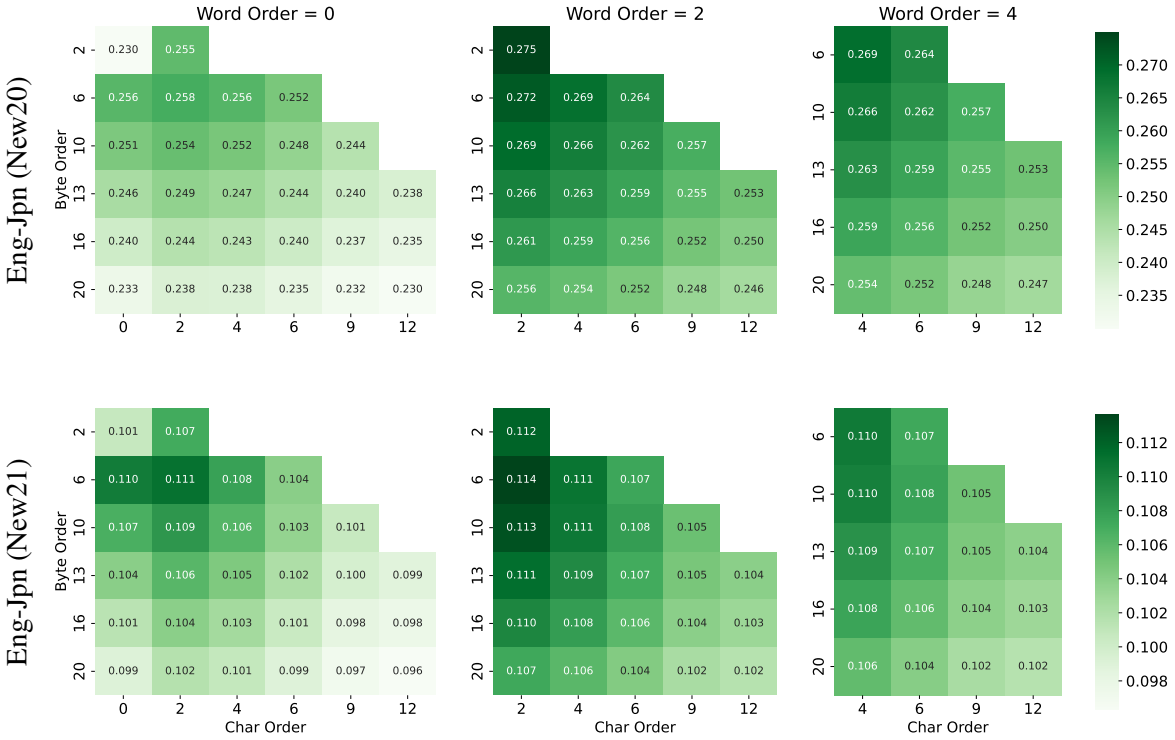


Figure 6: Pearson Correlation in relation between n-gram order of byte, character, and word on Japanese.

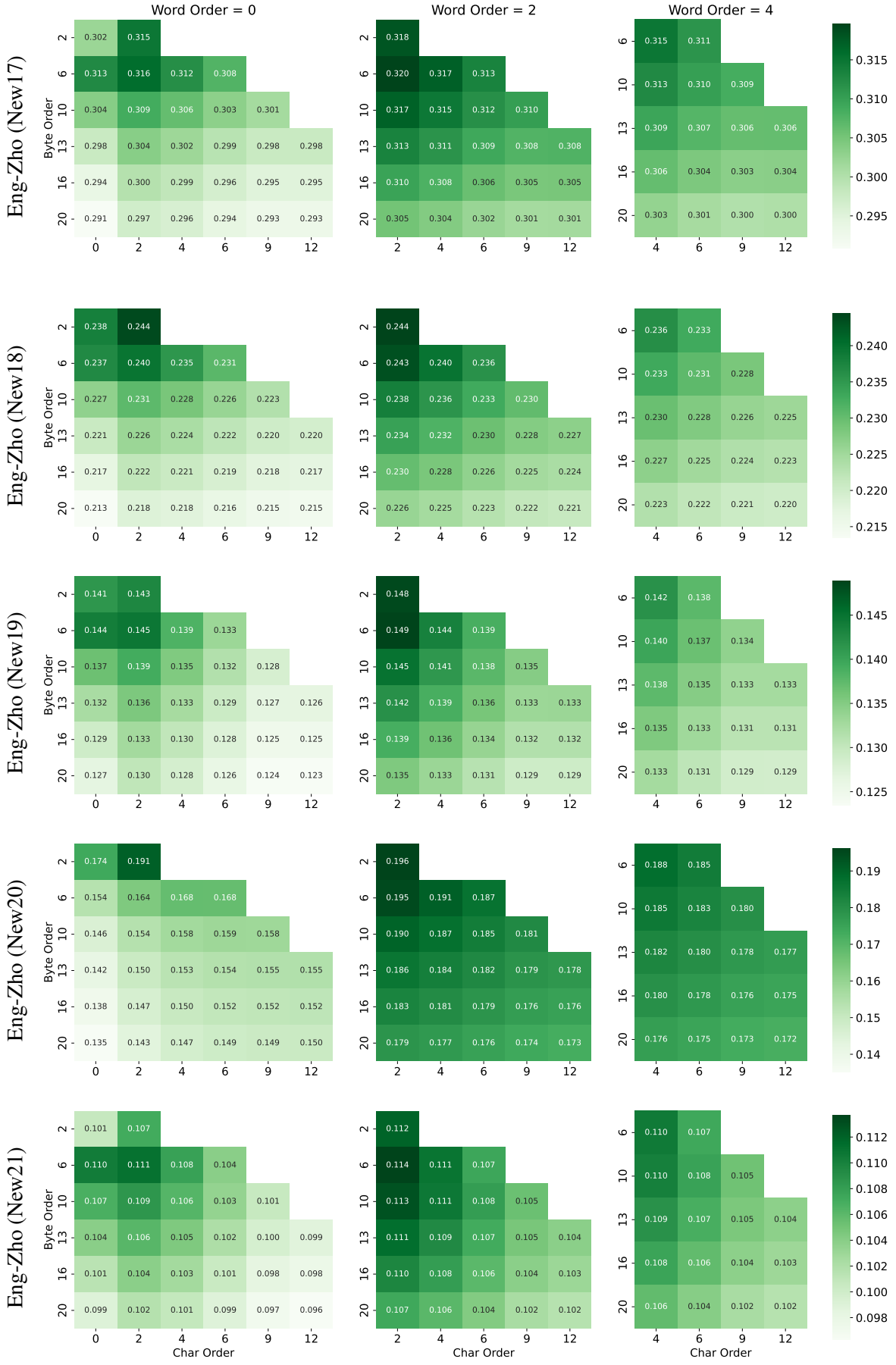


Figure 7: Pearson Correlation in relation between n-gram order of byte, character, and word on Chinese.

Quality Estimation and Post-Editing Using LLMs For Indic Languages: How Good Is It?

Anushka Singh[1,2] **Aarya Pakhale**[1,4]
Mitesh M. Khapra[1,2] **Raj Dabre**[1,2,3,5]

[1]Nilekani Centre at AI4Bharat [2]Indian Institute of Technology Madras, India
[3]National Institute of Information and Communications Technology, Kyoto, Japan
[4] Indian Institute of Technology Kharagpur, India
[5]Indian Institute of Technology Bombay, India

Abstract

Recently, there have been increasing efforts on Quality Estimation (QE) and Post-Editing (PE) using Large Language Models (LLMs) for Machine Translation (MT). However, the focus has mainly been on high resource languages and the approaches either rely on prompting or combining existing QE models with LLMs, instead of single end-to-end systems. In this paper, we investigate the efficacy of end-to-end QE and PE systems for low-resource languages taking 5 Indian languages as a use-case. We augment existing QE data containing multidimentional quality metric (MQM) error annotations with explanations of errors and PEs with the help of proprietary LLMs (GPT-4), following which we fine-tune Gemma-2-9B, an open-source multilingual LLM to perform QE and PE jointly. While our models attain QE capabilities competitive with or surpassing existing models in both reference-based and reference-free settings, we observe that they still struggle with PE. Further investigation reveals that this occurs because our models lack the ability to accurately identify fine-grained errors in the translation, despite being excellent indicators of overall quality. This opens up opportunities for research in end-to-end QE and PE for low-resource languages. The synthetic dataset and evaluation metrics are publicly accessible online.[1]

1 Introduction

The rapid advancements in Large Language Models (LLMs) (Achiam et al., 2023; Touvron et al., 2023; Riviere et al., 2024) have significantly impacted Machine Translation (MT) leading to state-of-the-art translation quality. This quality is usually measured at the corpus level using a variety of quality estimation (Zerva et al., 2024) metrics among which COMET (supervised) and GEMBA (prompting-based)(Kocmi and Federmann, 2023) are known to be the best. Specifically, COMET has spurred research into language-family specific versions of COMET like in the case of Indic langugaes (Sai B et al., 2023). Closely related is the problem of post-editing where once a poor quality translation has been detected, mistakes in translation need to be suitably fixed (Bhattacharyya et al., 2023).

Recently, Treviso et al. (2024) have shown that it is possible to take error annotations of COMET models and the power of synthetic explanations generated by GPT-4, to develop a system that can post-edit erroneous translations thereby improving translation quality. Their main focus was showing that error explanations in human understandable formats lead to improved post-edits by LLMs. On the other hand, Lu et al. (2025) have leveraged LLMs purely in prompting mode in multiple stages to first annotate errors, choose the most reliable ones, and then post-edit to improve translation quality. However, existing works have two major limitations: a. They do not focus on a singular end-to-end model which does error annotations, error explanations and post-editing in one go. b. They focus on high-resource languages, which makes it difficult to determine the impact on low-resource languages.

In this paper we attempt to fill this gap by focusing on English to Indian languages (En→X) directions – specifically for five Indian languages: Hindi, Gujarati, Marathi, Malayalam and Telugu, which are considered low-resource in the world of quality estimation and post editing. Given the low-resource setting, we ask a simple question: *How good is an all-purpose end-to-end error annotation, explanation and post-editing system for Indian languages in a low-resource setting?*. This leads to 3 specific research questions (RQs):

(RQ1): How well do Large Language Models per-

[1]https://github.com/AI4Bharat/QE-PE-MTEval.git

Proceedings of Machine Translation Summit XX Volume 1, pages 388–398, June 23-27, 2025

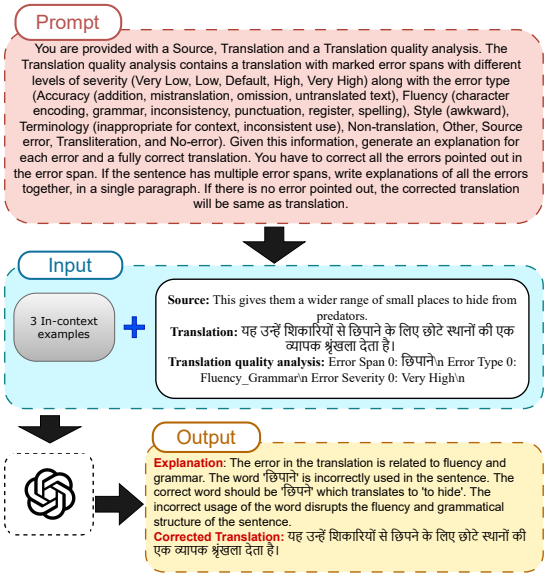


Figure 1: Overview of the approach used to generate synthetic post-edits and explanations. The figure illustrates the prompt design, input structure, and model-generated output. The prompt specifies how translation quality is analyzed, with error spans and severity levels guiding the generation of explanations and corrected translations

form in evaluating machine translation quality for Indian languages, considering both reference-based and reference-free scenario?

(RQ2): Do explanations of errors and error span detection by LLMs lead to demonstrable improvements in post-editing performance for Indian languages?

(RQ3): Does joint QE and PE, affect QE?

Taking motivation from (Treviso et al., 2024), we augment the Indic MT Evaluation dataset (Sai B et al., 2023) with synthetic explanations and post-edits (see Figure 1) and fine-tune GEMMA2(Riviere et al., 2024) to obtain a single model to generate error annotations (used for computing MQM scores for QE), error explanations and post-edits. On the positive side, we find that QE significantly surpasses all existing models like COMETKiwi, however, unlike previous works, we observe that error annotation and explanation does not often lead to higher translation quality after post-editing. Upon further investigation, we find that this mainly occurs because the limited amount of training data leads to models, which are good at evaluating overall translation quality, but are not always reliable at fine-grained quality estimation. Specifically, they tend to under-detect certain error categories or sometimes misclassify errors, leading to inconsistencies in post-editing corrections. This shows that we are still far away from using LLMs for fine-grained error annotation and use it for post-editing in low-resource settings. Our contributions are:

(i) State-of-the-art quality estimation models for 5 Indian languages in the En→X setting.

(ii) Augmented quality estimation dataset with error explanations and GPT4 post-edits.

(iii) A reality check that LLMs are still unreliable for fine-grained quality estimation and post-editing in low-resource settings.

2 Related Work

Research in the machine translation (MT) evaluation has evolved significantly, driven by the need for more accurate and interpretable metrics. Traditional MT evaluation metrics can be broadly classified into Reference-based and Reference-free approaches. Early metrics, such as BLEU (Papineni et al., 2002) and chrF (Popovic, 2017), primarily relied on lexical overlap between machine translations and human references, often failing to align well with human judgments.

More recent neural based metrics like, COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020) have shown stronger correlations with humans, but these metrics lack interpretability. These metrics have further improved with the introduction to models like XCOMET (Guerreiro et al., 2024) and COMETKiwi (in reference-free direction) (Rei et al., 2022, 2023) . However, XCOMET primarily detects error spans and their severity without classifying the specific type of error. We aim to explore whether LLMs can capture fine-grained translation errors by identifying their types alongside assessing severity, focusing on Indian languages.

In parallel, the exploration of Large Language Models (LLMs) for MT evaluation has gained momentum (Kocmi and Federmann, 2023; Xu et al., 2023), with research examining their effectiveness in assessing translation quality. While these approaches have been widely explored for high resource languages, their performance for Indian languages, which are notoriously resource poor for quality estimation, remains unexplored.

Additionally, research suggests that fine-grained error analysis and explanations can improve post-editing efficiency (Treviso et al., 2024; Lu et al., 2025). However, our findings indicate that such

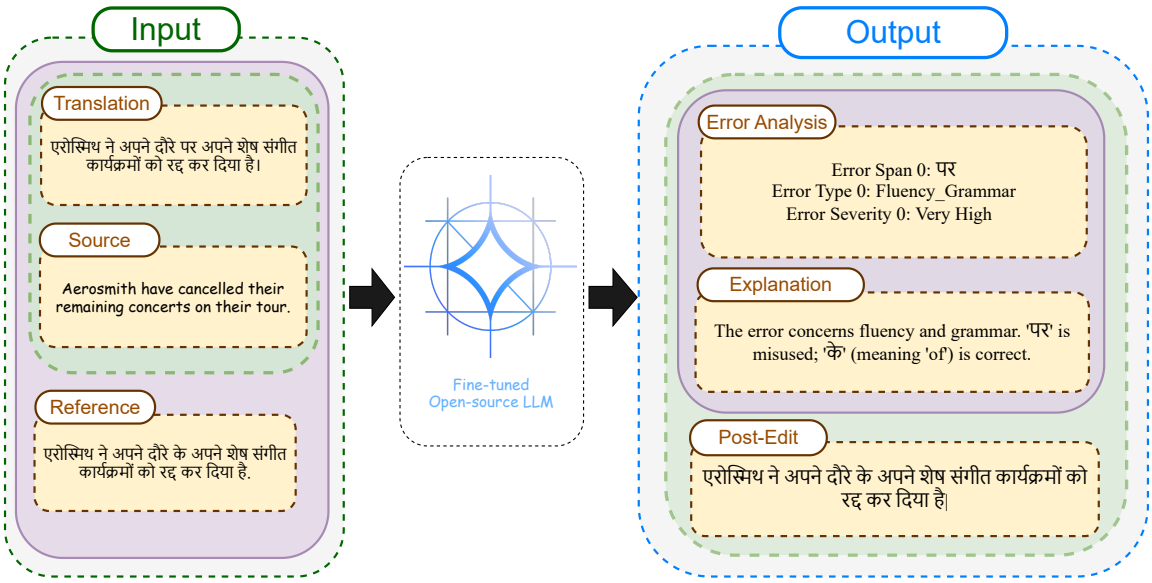


Figure 2: Overview of fine-tuned LLM models for translation quality assessment. The green box represents the reference-free setting, while the purple box represents the reference-based setting. Given an input consisting of a translation and source (with or without a reference and error analysis), we train models to generate one or more of the *error analysis* (fine-grained MQM style error annotations), *error explanations* and *post-edits* as applicable. Section 4.3 shows all possible model configurations we consider.

benefits may not necessarily extend to low-resource Indian languages, highlighting the need for further investigation into language-specific factors affecting post-editing and evaluation.

3 Methodology

Our approach leverages synthetic explanations and post-edits from LLMs followed by fine-tuning open-source LLMs to enhance a large language model's (LLM) ability to detect, explain, and correct machine translation errors in both reference-based and reference-free settings.

3.1 Error Explanations and Post-Edits

For the tasks of error analysis and post-editing, we generated synthetic explanations and post-edits using a proprietary API based model. Our approach, inspired by (Treviso et al., 2024) is shown in Figure 1. Our initial experiments with zero-shot prompting yielded suboptimal outputs, highlighting the need for more guided generation. To address this, we adopted a 3-shot prompting strategy, incorporating carefully selected in-context examples augmented with explanations and corrections.

The in-context examples were derived from expert annotations provided by bilingual linguists proficient in the target languages. Each linguist was presented with the source sentence, its machine translation, and pre-identified error spans, along with information on error type and severity. They were asked to provide detailed explanations for each error and generate a corresponding post-edited translation that reflects natural and fluent usage. Each expert annotated approximately 10 translation segments per language. From this pool, we manually selected three high-quality examples per language to serve as in-context demonstrations for API based model, enabling it to generate consistent and high-quality explanations and post-edits across the broader dataset.

3.2 Joint Quality Estimation and Post-Editing

Using the original QE data augmented with error explanations and post-edits, we fine-tune an open-source multilingual model in a variety of configurations. Figure 2 gives an overview and Section 4.3.2 details the training setups.

4 Experimental Setup

We now describe specifics of our experimental setup, namely datasets and languages, baselines, model configurations we tested, QE meta-evaluation and PE evaluation approaches.

Metric	Hindi		Malayalam		Marathi		Tamil		Gujarati		Average	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
COMET$_{MQM}$	0.441	0.597	0.405	0.516	0.365	0.490	0.498	0.654	0.426	0.487	0.427	0.549
Indic-COMET$_{MQM}$	0.479	0.656	0.441	0.557	0.394	0.538	0.523	0.677	0.473	0.552	0.462	0.596
Base-IndicBERT$_{MQM}$	0.438	0.638	0.443	0.517	0.370	0.512	0.437	0.576	0.487	0.582	0.435	0.565
XCOMET-XL	0.496	0.630	0.471	0.597	0.430	0.557	0.580	0.740	0.512	0.630	0.498	0.631
XCOMET-XXL	0.597	0.744	**0.642**	**0.696**	0.524	0.641	**0.602**	**0.747**	**0.610**	0.643	0.526	**0.694**
MetricX23-XL	0.419	0.401	0.457	0.427	0.388	0.406	0.465	0.396	0.452	0.449	0.436	0.416
MetricX23-XXL	0.439	0.333	0.417	0.391	0.476	0.421	0.323	0.478	0.323	0.533	0.438	0.422
MetricX24-XL	0.409	0.490	0.478	0.544	0.379	0.509	0.597	0.510	0.532	0.689	0.479	0.550
MetricX24-XXL	0.397	0.360	0.486	0.520	0.386	0.470	0.438	0.401	0.554	**0.720**	0.452	0.494
ErrSp	**0.776**	**0.778**	0.470	0.665	0.616	**0.657**	0.509	0.589	0.600	0.410	**0.594**	0.620
ErrSp-Exp	0.754	0.766	0.449	0.592	**0.637**	0.602	0.346	0.422	0.596	0.397	0.556	0.556

Table 1: Segment-level Pearson (ρ) and Kendall tau (τ) scores for evaluation models in the reference-based setting.

4.1 Languages and Dataset Agumentation

For our experiments, we employed the IndicMT Eval dataset(Sai B et al., 2023), which comprises 1,476 examples per language, covering Hindi, Marathi, Malayalam, Tamil, and Gujarati. The dataset was partitioned into training, validation, and test sets containing 1000, 200 and 276 examples, respectively, for each language.

To enrich the dataset with explanations and post-edits, we employed the GPT-4 API to generate synthetic explanations and post-edits using a 3-shot prompting strategy(refer Figure 1). Building upon existing prompt design (Treviso et al., 2024), we incorporated expert-annotated in-context examples to enhance the quality and relevance of the generated explanations and corrections.

While leveraging LLMs for synthetic data generation offers scalability, it also introduces challenges such as generic meta-phrases or contextually irrelevant content. To mitigate these, we iteratively refined prompts, curated in-context examples, and incorporated human verification steps. This meticulous process resulted in well-structured training and validation pairs tailored for error detection, explanation generation, and post-edit prediction.

Additionally, to gauge the quality and utility of the synthetic annotations, we conducted a human evaluation wherein annotators assessed 20 GPT-4-generated explanations per language. The feedback was largely positive, particularly for Hindi, Gujarati, and Marathi. These findings were further corroborated by COMET-22 score comparisons, which showed notable improvements in 76% of Hindi cases, 50% of Marathi, and 44% of Gujarati. Although Tamil (29%) and Malayalam (36%) saw more modest gains, they still reflect incremental improvements attributable to the synthetic data.

4.2 Implementation and Training

We fine-tuned the Gemma-2-9B (Riviere et al., 2024) model on a diverse set of machine translation evaluation tasks, as shown in Figure 2. We initially experimented with fine-tuning LLaMA-3 (Touvron et al., 2023) models; however, their performance was suboptimal compared to Gemma-2, and hence we focused only on the latter. Fine-tuning was conducted with LoRA with a rank of 2 and an alpha value of 16 to optimize memory efficiency while maintaining model performance. For training we used a batch size of 8, a learning rate of 1.5e-4, and BF16 precision. Training was conducted using the open-instruct library[2].

4.3 Models Compared

We describe baselines followed by our various model configurations we tested.

4.3.1 Baselines

All existing baselines we consider only have the capability to do QE and we compare them with the QE capabilities of models we train. We compared our QE results against COMET(MQM) (Rei et al., 2020), IndicCOMET and its variants (Sai B et al., 2023; Singh et al., 2024), MetricX23 (Juraska et al., 2023), MetricX24 (Juraska et al., 2024), XCOMET (in a reference-based setting), and COMETKiwi (for a reference-free setting).

4.3.2 Our Models

We have reference-based models for QE and error explanation and reference-free models for QE, error explanation, and PE. Detailed in Appendix A

Reference-based QE Models These take in source, translation and a reference and produce:

1. **ErrSp:** Error Annotations (error spans).

[2] https://github.com/allenai/open-instruct

Metric	Hindi		Malayalam		Marathi		Tamil		Gujarati		Average	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
COMET_QE$_{MQM}$	0.487	0.651	0.354	0.457	0.302	0.416	0.485	0.650	0.359	0.370	0.397	0.509
IndicCOMET$_{MQM}$	0.507	0.675	0.424	0.507	0.349	0.470	**0.526**	**0.680**	0.434	0.428	0.448	0.552
Base-IndicBERT$_{MQM}$	0.439	0.632	0.409	0.520	0.362	0.479	0.476	0.596	0.445	0.547	0.426	0.555
COMET_Kiwi	0.542	0.634	0.458	0.480	0.392	0.475	0.482	0.393	0.494	0.681	0.474	0.533
COMET_Kiwi-XL	0.521	0.586	0.448	0.457	0.405	0.480	0.458	0.287	0.498	0.581	0.466	0.478
COMET_Kiwi-XXL	0.528	0.646	0.448	0.501	0.415	0.526	0.473	0.479	0.451	0.605	0.463	0.551
MetricX23-XL	0.464	0.455	0.423	0.285	0.371	0.300	0.447	0.197	0.443	0.503	0.430	0.348
MetricX23-XXL	0.550	0.417	0.484	0.334	0.424	0.369	0.499	0.241	0.538	0.600	0.499	0.392
MetricX24-XL	0.424	0.593	0.419	0.492	0.326	0.443	0.465	0.486	0.482	0.650	0.423	0.533
MetricX24-XXL	0.461	0.581	0.454	0.501	0.386	0.459	0.399	0.435	0.517	**0.717**	0.443	0.539
ErrSp	0.779	**0.777**	0.641	0.429	0.619	0.634	0.438	0.536	**0.611**	0.403	0.618	0.556
ErrSp-Exp	0.726	0.731	0.594	0.434	**0.621**	**0.644**	0.456	0.374	0.575	0.368	0.594	0.510
ErrSp-Exp-PE	0.754	0.765	0.656	0.457	0.588	0.621	0.370	0.479	0.582	0.374	0.590	0.539
ErrSp-Exp-PE*gpt*	0.753	0.763	0.569	0.452	0.567	0.592	0.443	0.361	0.541	0.343	0.575	0.502
ErrSp-PE	0.753	0.742	**0.697**	**0.560**	0.615	0.642	0.473	0.561	0.604	0.412	**0.628**	**0.583**
ErrSp-PE*gpt*	**0.783**	0.773	0.672	0.506	0.586	0.612	0.455	0.523	0.584	0.368	0.616	0.556

Table 2: Segment-level Pearson (ρ) and Kendall tau (τ) scores for evaluation models in the referenceless setting.

2. **ErrSp-Exp:** 1 + human readable explanations (henceforth explanations).

Reference-free QE and PE Models These take in only source and translation and produce:

1. **ErrSp:** Error Annotations (error spans).
2. **ErrSp-Exp:** 1 + explanations.
3. **PE:** Post-edits with the original reference was used as the post-edit during training.
4. **PE$_{gpt}$:** Post-edits with the GPT generated correction as the post-edit during training.
5. **ErrSp-Exp-PE:** 2+3
6. **ErrSp-Exp-PE$_{gpt}$:** 2+4
7. **ErrSp-PE:** 1+3
8. **ErrSp-PE$_{gpt}$:** 1+4

Additionally, we trained some control models specifically for the purposes of PE, to determine if PE quality improves when the correct error spans are supplied to the model as a part of the prompt (*ip*). To this end, we take the correct error spans as inputs along with the source and translation as a part of the model prompt when training.

9. **ErrSp-ip-PE:** Analogous to 7.
10. **ErrSp-ip-PE$_{gpt}$:** Analogous to 8.
11. **ErrSp-ip-Exp-PE:** Analogous to 5.
12. **ErrSp-ip-Exp-PE$_{gpt}$:** Analogous to 6.

4.4 QE and PE Evaluation

To meta-evaluate the QE capabilities of models, we follow Rei et al. (2020) and compute Pearson and KendallTau correlations of MQM scores computed using predicted MQM error spans against those done by humans. For PE, we compute chrF (Popovic, 2017) and COMET-22 scores of the post-edit generated by the model against the human written reference.

5 Result

In this section, we present the evaluation results of our LLM-based approach for MT quality assessment of Indian languages, addressing the research questions outlined in Section 1. Section 5.1 addresses **RQ1** by evaluating the performance of our models under both reference-based and reference-free settings, comparing them against state-of-the-art MT evaluation systems. Section 5.2 focuses on **RQ2**, investigating whether error annotations and explanations enhance post-editing performance. Additionally, throughout both sections, we explore **RQ3**, analyzing whether joint quality estimation (QE) and post-editing (PE) influence QE performance. By structuring our results around these questions, we provide a comprehensive assessment of LLM capabilities for low-resource MT evaluation.

5.1 LLM-Based MT Evaluation for Indian Languages

Table 1 presents the results of reference-based MT evaluation. Our LLM-based approach achieves competitive performance, comparable to the significantly larger XCOMET-XXL (10.7B) model. Notably, unlike XCOMET-XXL, our method identifies error spans with greater diversity in both category and severity (refer Table 6 for details). Our system demonstrates strong performance for Hindi and Marathi, but we observe comparatively lower

Metric	Hin	Mal	Mar	Tam	Guj
pre-edit	**48.89** / 0.737	47.67 / 0.839	48.47 / 0.729	48.33 / 0.850	50.96 / 0.851
PE	45.24 / 0.733	45.06 / 0.842	42.71 / 0.703	45.80 / 0.854	44.99 / 0.851
PE*gpt*	48.86 / 0.733	48.46 / 0.842	48.89 / 0.703	49.12 /0.854	51.59 / 0.851
ErrSp-PE	45.16 / 0.738	45.69 / 0.840	43.69 / 0.711	47.41 / 0.863	47.05 / **0.859**
ErrSp-PE*gpt*	48.66 / **0.743**	48.03 / 0.832	48.75 / 0.734	48.60 / 0.843	51.17 / 0.846
ErrSp-Exp-PE	47.12 / 0.665	43.54 / 0.736	43.63 / 0.684	44.97 / 0.669	47.45 / 0.761
ErrSp-Exp-PE*gpt*	46.69 / 0.673	44.78 / 0.725	46.12 / 0.698	44.06 / 0.707	47.47 / 0.745
ErrSp-ip-Exp-PE	46.96 / 0.731	45.26 / 0.838	43.53 / 0.717	47.69 / 0.841	45.17 / 0.843
ErrSp-ip-Exp-PE*gpt*	47.43 / 0.714	46.32 / 0.815	48.82 / 0.730	48.40 / 0.815	49.52 / 0.831
ErrSp-ip-PE	44.61 / 0.730	44.85 / 0.837	43.50 / 0.707	46.06 / 0.856	46.90 / 0.855
ErrSp-ip-PE*gpt*	48.70 / **0.743**	**48.92** / **0.845**	**49.00** / **0.737**	**49.99** / **0.858**	**51.71** / 0.854

Table 3: ChrF and COMET scores of model-suggested post-edits vs. reference. Scores are in X/Y format, where X is ChrF and Y is COMET. The "pre-edit" row shows ChrF and COMET scores for MT output vs. reference.

performance for Gujarati and Tamil. This discrepancy suggests language-specific challenges, which require further investigation.

The reference-free evaluation results in Table 2 highlight that our model achieves state-of-the-art performance. Specifically, our model ranks second-best when only predicting error spans but outperforms all models when tasked with both error span detection and post-editing. This underscores the effectiveness of LLMs in evaluating MT quality, particularly when integrating error correction. Consistent with our reference-based findings, the strongest performance is observed for Devanagari-script languages (Hindi and Marathi), reinforcing the notion that script and linguistic features play a crucial role in quality estimation. We also observed that our model got a relatively lower Pearson score; the reason can be the non-linear relationship between model predicted scores and actual MQM scores, the presence of clustered values around certain score ranges (e.g., 0.6, 0.8, and 1.0), and the skewed distribution, which weakens Pearson ability to capture a strong linear correlation despite maintaining a high rank correlation (KendallTau).

5.2 Impact of Error Analysis on Post-Editing

In this section, we analyze the impact of error analysis on post-editing, with a particular focus on **RQ3**, which examines whether joint quality estimation (QE) and post-editing (PE) influence QE performance. Table 3 presents ChrF++ and COMET scores for both original machine translations (pre-edits) and their best post-edited versions. Contrary to prior work suggesting that error explanations significantly improve post-editing quality (Treviso et al., 2024), our results show only marginal gains across Indian languages. Interestingly, while error detection leads to notable improvements in reference-free QE (as shown in Section 5.1), these gains do not consistently carry over to post-editing. The highest ChrF++ and COMET scores are observed when error annotations are available, yet the improvements remain modest, underscoring the limitations of LLM-based post-editing in low-resource settings. Our findings suggest that joint modeling of QE and PE does not consistently enhance QE performance. Although the best results are achieved when combining error analysis with post-editing, the addition of explanations does not yield further benefits. One potential reason for this can be the scarcity of high-quality training data. In contrast to high-resource languages, where fine-grained error analysis and explanations can drive significant improvements, LLMs struggle to generate precise, actionable feedback for low-resource languages. These results indicate that while LLMs show promise in overall MT quality estimation, they remain less reliable for fine-grained quality assessment and post-editing in low-resource scenarios.

6 Conclusion

Our study investigates the role of Large Language Models (LLMs) in machine translation (MT) evaluation for Indian languages, addressing key challenges in fine-grained quality estimation (QE) and post-editing (PE). We leveraged synthetic error explanations and post-edits from GPT-4 and fine-tuned the GEMMA-2-9B model in a variety of settings for reference-based QE and reference-free QE and PE. In reference-based settings we got comparable if not slightly better QE performance against existing strong baselines. On the other hand, in reference-free settings we obtained significantly

improved QE performance. However in the case of PE, contrary to previous works in high-resource settings, involving error detection and explanation in the PE framework does not lead to improved post-edited translations. The explanation for this is in the poor fine-grained error detection capabilities of our fine-tuned models due to low-resource settings. This indicates a dire situation but opens avenues for future research on joint QE and PE for low-resource languages.

7 Limitations

This study examined LLM performance on a selection of Indian languages. Future research should broaden this scope to encompass a more diverse set, particularly low-resource languages. Furthermore, even with fine-tuning, LLM post-editing performance for Indian languages requires improvement. To this end, better strategies for low-resource post-editing need to be studied. Another limitation of this work is the limited amount of synthetic data created which should also be a future topic of investigation.

8 Sustainability Statement

Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.45 kgCO_2eq/kWh. A cumulative of 48 hours of computation was performed on hardware of type A100 PCIe 40GB (TDP of 250W). Total emissions are estimated to be 5.4 kgCO_2eq of which 0 percents were directly offset. Given the low-resource nature of our work, we do not expect our work to have any large negative environmental impact.

Estimations were conducted using the Machine-Learning Impact calculator presented in (Lacoste et al., 2019).

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, Haiming 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, Isabella 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, Johannes 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, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Adeola 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 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 D. 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 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, Qiming 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.

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, 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.

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 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.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

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.

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 Popovic. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 612–618. 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, 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.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stańczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Boxi Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Christoper A. Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozi'nska, D. Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Pluci'nska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost R. van Amersfoort, Josh Gordon, Josh Lipschultz, Joshua Newlan, Junsong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Mil-

lican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, L. Sifre, Lena Heuermann, Leti cia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Gorner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Peng chong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, S'ebastien M. R. Arnold, Se bastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomás Kociský, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeffrey Dean, Demis Hassabis, Koray Kavukcuoglu, Clément Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. 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, Bangkok, Thailand. 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, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Marcos V Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan Van Stigt, and Andre Martins. 2024. xTower: A multilingual LLM for explaining and correcting translation errors. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15222–15239, Miami, Florida, USA. 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.

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.

Model Name	Inputs Provided	Outputs Expected
	Reference-Based	
ErrSp	Source, Translation, Reference	Error Spans
ErrSp-Exp	Source, Translation, Reference	Error Spans + Explanations
ErrSp-ip-Exp	Source, Translation, Reference, Error Spans	Explanations
	Reference-Free	
ErrSp	Source, Translation	Error Spans
ErrSp-Exp	Source, Translation	Error Spans + Explanations
ErrSp-Exp-PE	Source, Translation	Error Spans + Explanations + Post-Edits
ErrSp-ip-Exp	Source, Translation, Error Spans	Explanations
ErrSp-ip-Exp-PE	Source, Translation, Error Spans	Explanations + Post-Edits
ErrSp-ip-PE	Source, Translation, Error Spans	Post-Edits
ErrSp-PE	Source, Translation	Error Spans + Post-Edits
PE	Source, Translation	Post-Edits

Table 4: Overview of GEMMA fine-tuning tasks under reference-based and reference-free settings. Each task is defined by the specific inputs provided and the expected outputs the model learns to generate.

Metric	Hin	Mal	Mar	Tam	Guj
Err_Sp Exp	59.46	46.18	55.03	46.78	52.05
Err_Sp Exp PE	58.83	55.11	46.10	49.31	52.97
Err_Sp_Exp PE-gpt	59.26	49.37	41.74	43.71	47.87
Err_Sp_ip Exp	70.17	61.63	55.70	65.47	61.65
Err_Sp_ip Exp PE	70.20	60.94	55.29	64.96	60.81
Err_Sp_ip Exp PE-gpt	70.46	61.67	56.35	65.38	61.47

Table 5: chrF scores of model-suggested explanation vs. GPT generated explanation

A Training Data Preparation

To fine-tune GEMMA-9B for translation quality estimation and post-editing tasks, we constructed a diverse set of input-output training pairs using synthetic error explanations and post-edits. The model was trained under two major settings: *reference-based* (using human reference translations) and *reference-free* (using only the source and machine translation). Table 4 summarizes the task variants explored under each setting.

Figure 2 shows an example prompt for the ErrSp-Exp-PE task in the reference-free setting. Other task prompts follow similar structures, differing in the presence or absence of reference translations, error spans, or expected outputs (e.g., explanations, corrections).

For reference-free training, we experimented with two post-edit supervision strategies: one using GPT-4 generated outputs ($\texttt{PE}_{\texttt{gpt}}$), and the other using human references (PE). This comparison helps evaluate the reliability of synthetic supervision in low-resource scenarios.

Error Category		Explanation
Accuracy	Addition	Translation includes information not present in the source.
	Omission	Translation is missing content from the source.
	Mistranslation	Translation does not accurately represent the source.
	Untranslated text	Source text has been left untranslated
Fluency	Spelling	Incorrect spelling or capitalization.
	Grammar	Problems with grammar, other than orthography.
	Register	Wrong grammatical register (eg, inappropriately informal pronouns).
	Character Encoding	Characters are garbled due to incorrect encoding. Example: Sink ->$ink
Terminology Inappropriate		Terminology is non-standard or does not fit context.
Style Awkward		The style of the text does not feel very apt. (Example: 1. The source sentence feels formal like in a newspaper, but the translation doesn't. 2. Sentences are correct, but simply too long, etc..)
Transliteration		If it transliterates instead of translating words/ phrases, where it should not.
Other		Any other issues.
Source Error		An error in the source.
Non Translation		Impossible to reliably characterize the 5 most severe errors.

Table 6: This table outlines the error categories our models are capable of detecting in machine translation outputs. It includes a comprehensive list of common translation errors, ranging from accuracy issues like additions and omissions to fluency problems such as spelling and grammar mistakes. The categorization is adapted from previous work IndicMT-eval(Sai B et al., 2023)

Research – Translators and Users

Revisiting Post-Editing for English-Chinese Machine Translation

Hari Venkatesan
University of Macau
hariv@um.edu.mo

Abstract

Given the rapid strides in quality made by automated translation since the advent of Neural Machine Translation, questions regarding the need and role of Post-Editing (PE) may need revisiting. This paper discusses this in light of a survey of opinions from two cohorts of post-graduate students of translation. The responses indicate that the role of PE may need further elaboration in terms of aspects such as grammar, lexis and style, with lexis and style being the main sites requiring human intervention. Also, contrary to expectations, responses generally show marked hesitation in considering quasi-texts as final without PE even in case of disposable texts. The discussion here pertains to English-Chinese translation, but may resonate with other language pairs as well.

1 Introduction

Post-Editing or simply editing as a phenomenon may have existed ever since writing and the need to revise came into existence. However, the concept this paper is concerned with is Machine Translation Post-Editing (MTPE) where "...the task of the post-editor is to edit, modify and/or correct pre-translated text that has been processed by an MT system from a source language into (a) target language(s)." (Allen, 2003, p. 297)

The manner in which post-editing was conducted evolved from the paper and pencil work to editing on a word-processor and eventually through interactive software systems (Hutchins & Somers, 1997, p. 153).

MTPE further came to be classified broadly in terms of the extent of editing and targeted quality into minimal post-editing (for inbound purposes) and maximal post-editing (for publication and outbound purposes) (Allen, 2003, pp. 301–303).

Given the relatively lower quality of unedited or quasi-texts (Allen, 2003, p. 298) produced by MT, early discussions considered using unedited texts for "gisting" or as a pre-translation for screening (Allen, 2003, p. 303; Hutchins & Somers, 1997, p. 157). Later, international standards were evolved for MTPE such as the International Standards Organization's (ISO 18587:2017 Translation Services — Post-Editing of Machine Translation Output — Requirements, 2017) that classifies MTPE into Light Post-Editing and Full Post-Editing. According to this standard, Light Post-Editing is a "process of post-editing (3.1.4) to obtain a merely comprehensible text without any attempt to produce a product comparable to a product obtained by human translation (3.4.3)"; Full post-editing on the other hand refers to "process of post-editing (3.1.4) to obtain a product comparable to a product obtained by human translation (3.4.3)".

The definition of Light Post-Editing here is akin to what Allen terms "Rapid Post-Editing" where "a strictly minimal number of corrections on documents that usually contain perishable information" (Allen, 2003, p. 302). It must be mentioned here that this standard was created in 2017 and it is currently under review. Detailed guidelines for MTPE are also provided by the Translation and Automation User Society (TAUS), which makes a similar distinction between light and full PE but suggests creating "a clear matrix of post-editing productivity, quality, turnaround time and pricing discount expectation" based on a detailed analysis (Massardo et al., 2016, p. 12). TAUS guidelines also provide for the possibility of "Good Enough" quality that involve ensuring

Proceedings of Machine Translation Summit XX Volume 1, pages 399–406, June 23-27, 2025

semantically correct translation and making no stylistic changes or changes intend to enhance naturalness (Ibid, p. 17). In addition to these many studies have proposed models to arrive at PE decisions or achieve quality goals based on purpose and nature of text being translated (Nitzke et al., 2024; Rico Pérez, 2024; Venkatesan, 2022).

However, given the rapid developments in MT, particularly the emergence of widely available NMT starting around 2016 and more recently Large Language Model based generative AI such as ChatGPT and DeepSeek, the quality of output achieved by MT has vastly increased. It is therefore important to ask which aspects of MTPE, if at all, remain relevant and whether translators perceive a clear distinction between levels of PE that may be required.

2 Post-Editing in the era of NMT and AI

With specific reference to English-Chinese translation, as early as 2018 there were claims of MT having achieved parity with Human Translation (HT) in domains such as news translation (Hassan et al., 2018), though evidence for human translation being superior were also presented (Läubli et al., 2018). The quality of raw output from Machine Translation has increased across domains and recent studies have shown that translations produced by MTPE "were more accurate than the outputs from HT [Human Translation] both for STs of high and low complexity" (Jia & Sun, 2023, p. 963), even though the authors do not report a strong co-relation between perceived and actual difficulty measurement when comparing MTPE and HT. A 2021 study involving Chinese translator trainees also demonstrated increased speed and reduced effort on the part of translators (Wang et al., 2021) when using MTPE. As previously mentioned, even if quasi-texts produced by MT are not considered entirely free of errors, in the interests of efficiency and particularly for general everyday communication, there have always been suggestions that raw MT output may be suitable for gisting or may simply undergo light post-editing to eliminate critical errors. Given the advanced in MT quality today, it may be assumed that the possibility of using MT without editing should be higher, at least for some purposes. A recent study that revisited definitions of light post-editing and full post-editing suggested that these definitions may no longer be valid and instead advocates redefining MTPE guidelines based on an ecosystem incorporating all aspects that influence a translator's decision making (Rico Pérez, 2024). The question now is to what extent have the strides made by MT resulted in reduced necessity for PE? Given the improvements in quality, does the distinction between light and full PE continue to hold good? In the following we discuss how postgraduate trainee translators perceive the quality produced by MT in general and the nature and role of PE in particular. For this purpose, a survey of opinions was conducted with two successive cohorts (academic year 2022/23 and 2023/24) of post-graduate students of translation as respondents.

3 Survey

The survey employed a questionnaire (see Appendix A) with responses graded according to the Likert scale. In the data presented below the responses are assigned scores from 0 to 4 (0 for 'Strongly Disagree', 1 for 'Disagree', 2 for 'Neutral', 3 for 'Agree' and 4 for 'Strongly Agree). The survey was tested and adjusted for clarity. In terms of reliability, using the survey data on SPSS a Cronbach Alpha score of 0.860 and 0.872 was obtained for 2023 and 2024 respectively, which suggest good internal consistency.

The data presented below shows responses from two cohorts of students from the Master of Arts in Translation Studies programme at the University of Macau. Each cohort is made of 25-30 students, of whom 20 from 2024 and 21 from 2023 responded to the survey. A majority of students admitted to the programme come from different parts of mainland China, while roughly a quarter come from Macao SAR. All students go through a rigorous assessment of language proficiency and preparation before admission. All respondents reported falling under the 20-30 years age group with 10-15 years of formal education in English and 15-20 years of formal education in Chinese on average. All students have Chinese (Putonghua or Cantonese) as their primary language and English as their second or acquired language. The students attended a compulsory course titled "Translation Technology" that discussed definitions of MTPE in detail and also trained students to carry out light PE and full PE. They were asked to respond to survey questions based on their experience of post-editing

translations both from Chinese to English and English to Chinese for different genres of writing.

The purpose of the survey was to understand how postgraduate students of translation perceived the role of MTPE in terms of need and scope. The purpose behind repeating the survey over two years was to see if there were significant changes in the attitudes and perception conveyed given the normalization of MT use that is expected to occur with time. The survey questionnaire was divided into three parts: Questions regarding efficiency and quality of MTPE (1a-1d), questions regarding types of MTPE required (meaning aspects that most require PE, 2a-2l), and necessity of PE for MT produced for different purposes (3a-3f). The mean scores for each question are shown in figures 1 and 2 below:

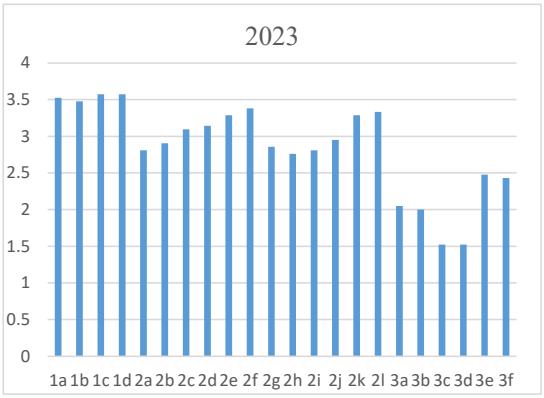


Figure 1: Responses from 2023

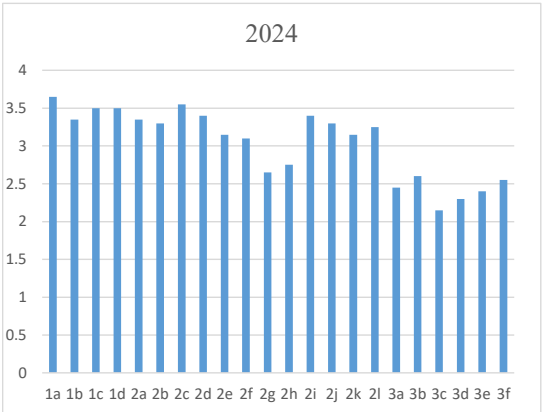


Figure 2: Responses from 2024

3.1 Results

The first part of the questionnaire (1a-1d) asks if MT helped increase efficiency (1a, 1c) and quality (1b, 1d) in case of E>C and C>E translation respectively. In case of both cohorts, responses ranged between 'Agree' and 'Strongly Agree' in case of efficiency, and also in the case of quality, albeit with slightly lower scores for quality in E>C translation. The standard deviation in responses is shown in Appendix B and C. The deviation in 2023 is largest in case of 1d (0.79) and remained under 0.5 in case of 1b and 1c and a little over it (0.58) in case of 1a.

The first six questions of the second part (2a-2f) ask if PE is necessary to correct grammatical, lexical and stylistic errors in both directions. Respondents from 2023 on average seemed to suggest that this was more necessary in case of C>E translation in each case with responses ranging from "Neutral" to "Agree" in case of grammatical errors, slightly over "Agree" in case of lexical errors and between "Agree" and "Strongly Agree" in case of stylistic errors, indicating relatively high confidence in grammar produced by MT. Respondents from 2024, on the other hand, similarly rated the need to edit for lexical errors higher than grammatical errors, but unlike those from 2023 considered stylistic errors as least important when it comes to PE. Also, unlike 2023 there is a slight reversal observed with E>C perceived as more in need for PE in all three cases. The standard deviation observed in responses to these questions was highest (1.10) in case of 2f and high for 2d and 2e (0.90) in case of 2023. High standard deviation was observed in responses to 2b (1.1) and 2e and 2f (0.79 and 0.7 respectively) in case of 2024. This indicates more relative divergence on the question regarding PE for stylistic errors in C>E translation. In case of 2024 high divergence is observed in responses to the question whether PE is required to correct grammatical errors in C>E translation.

The following six questions (2g-2l) were regarding the extent of PE required to make a text publishable (2g-2j) and whether there was a significant difference between light and full PE (2k-2l). Respondents from both years seemed to fall between "Neutral" and "Agree" to the suggestion that light post-editing was sufficient to make texts publishable in case of MT in either direction. To the suggestion that full post-editing was essential to make texts publishable, both years were relatively more affirmative with 2023 still falling between "Neutral" and "Agree" but close to "Agree", while

2024 fell between "Agree" and "Strongly Agree". As for the question of whether there was a significant difference between light and full post-editing both years had average responses situated between "Agree" and "Strongly Agree", albeit closer to "Agree". The standard deviation for this group of questions was relatively high (0.62 to 1.10) in case of 2023 and 2024 (0.53-1.23). In both years highest divergence (1.10 and 1.23) is noted in responses to question 2g whether light PE is sufficient to make raw MT (C>E) publishable, with responses ranging from "Disagree" to "Agree". High deviation is also noted for 2i-2k, with only 2l showing relatively low deviation.

The third part of questionnaire juxtaposes the need for PE with end use (inbound, outbound and disposable). On the suggestion that inbound MT need not be post-edited, responses from 2023 were "Neutral" while those from 2024 ranged between "Neutral" and "Agree" with 3a (E>C) scoring marginally higher in 2023 and 3b (C>E) in 2024. When the question was changed to being about outbound translation (3c-3d) responses from 2023 ranged between "Disagree" and "Neutral", reaching about the mid-point on average while those from 2024 remained between "Neutral" and "Agree", albeit with lower averaged than the previous set of question regarding inbound translation but with C>E scoring higher. For the last two questions suggesting that no PE was needed in case of disposable texts responses from both 2023 and 2024 ranged between "Neutral" and "Agree", though responses almost touched the mid-point between "Neutral" and "Agree" in case of 2023 while remaining marginally short in 2024 in case of E>C and marginally over the mid-point in case of C>E. The standard deviation observed in case of these responses was high in case of 2024 (1.15 to 1.32) and high except for the first three questions (3a-3c) in case of 2023. The last question (3f) that suggested that C>E MT of disposable texts need not be post edited showed the widest divergence of 1.01 and 1.32 in case of 2023 and 2024 respectively, implying responses ranging from "Disagree" to "Agree". It is apparent that there was generally a wide divergence in opinions on the suggestion of doing away with PE for raw MT output.

3.2 Discussion

The results of the survey are intriguing as they show variations, albeit minor, even when it comes to the direction of translation. For instance, respondents seem relatively more affirmative of MT (without PE) in the C>E direction as shown in responses to 1c, 1d (with the exception of 1a which received the highest score in 2024), 2b, 2d, 2f, 2j, 2l in 2023, while this reverses with 2024 rating E>C MT higher in 2a, 2c, 2e, 2i. In case of 3a-3f C>E shows relatively higher averages in 2024, while responses from 2023 for this group remain largely the same, with E>C marginally highest for 3a and 3e. While the variations are minor, they may indicate varying levels of confidence in either language and differences in the ability to spot errors in quasi texts.

All respondents seem to agree that MT+PE increases both efficiency and quality in both directions, this shows general acknowledgement and recognition of current quality achievable by MT. The suggestion that MT+PE increases quality in E>C translation shows slightly lower averages, which may be understandable given that the respondents have Chinese as their first language.

In case of respondents from 2023, PE seems to be seen as necessary mostly for stylistic changes, while PE for lexical and grammatical errors stood lower, in that order. PE for Grammatical errors also seemed to rank low in importance also as the responses ranged between "Neutral" and "Agree" unlike those for lexical and stylistic errors that ranged between "Agree" and "Strongly Agree". Furthermore, in each case the need for PE seems to be felt more in the case of C>E translation. Responses from 2024 on the other hand show highest scores in case of need of PE for lexical errors, while grammatical errors and stylistic errors followed. Again unlike 2023, PE for E>C translation received slightly higher scores in each case. However, the need for PE in all three cases ranged between "Agree" and "Strongly Agree". The relatively low score in both years for the need for PE to correct grammatical errors seems to endorse the maturing of MT in terms of being error free at the grammatical level. Interestingly, lexical and stylistic errors seem to be seen as a more important site of errors necessitating PE. This result resonates with studies that have found that MT may

sometimes leave content untranslated or mistranslated (Goto & Tanaka, 2017).

On the question whether light PE was sufficient to make MT publishable or full PE was necessary (2f-2j), there was only a slight difference in 2023 with responses ranging between "Neutral" and "Agree". However, in case of 2024 responses ranged between "Neutral" and "Agree" on the suggestion that light PE was sufficient, and "Agree" and "Strongly Agree" on the suggestion that full PE was necessary. Both years also showed responses between "Agree" and "Strongly Agree" to the suggestion that time taken in light and full PE in either direction was significantly different. In summary, there is both agreement and reservation expressed to the idea that light PE may be sufficient to make MT publishable. In both years, this question (2g-2f) shows a relatively large standard deviation, suggesting less convergence in perception. The deviation was slightly lower, between 0.7 and 0.95 on the suggestion that full PE was essential. Finally, all seemed to be more in accord with the suggestion that time taken for light and full PE was significantly different. What is interesting is that average scores for the first two sets of questions (2g-2j, on light and full PE) were nearly identical in 2023, while there was a clear difference in 2024 with need for full PE scoring higher. There in some ambivalence in responses as there is endorsement of the quality of raw MT output and also the possibility that light PE may be sufficient to make texts publishable, but all respondents seem to agree that time and effort in light and full PE are significantly different and seem to suggest that full editing is essential to make a text publishable. Combining this with responses from the previous set of questions, it seems to suggest that while grammatical concerns may not be as serious as before, lexical and stylistic errors continue to require PE, which might fall under the category of full PE.

The third part of the survey makes more direct suggestions to examine what kind of texts produced by MT may summarily do away with PE. Responses from 2023 showed the highest affirmation towards doing away with PE in case of disposable texts (3e-3f), followed by inbound texts (3a-2b) and outbound texts (3c-3d) largely along expected lines. The suggestion regarding no PE for outbound texts adds to questions 2g-2j in a different way to test limits that respondents may be comfortable with. While responses ranged between "Neutral" and "Agree" in 2023 and "Neutral and "Agree" for light-PE and "Agree" and "Strongly Agree" for full-PE in 2024, responses to 3c-3d (no PE for outbound texts) range from "Disagree" to "Neutral" in 2023 and slight above "Neutral" but less than half way between "Neutral" and "Agree" in case of 2024. This shows definite discomfort with the idea of no PE for outbound texts. In case of no PE for inbound texts (3a-3b) responses were on average "Neutral" in case of 2023 and slightly between "Neutral" and "Agree" in case of 2024. It is interesting to note that while endorsing the largely error-free nature of MT (at least in terms of grammar) respondents from both years are still not very confident about doing away with PE even in case of inbound texts. It is only with disposable texts (3e-3f) that the responses range between "Neutral" and "Agree", again not emphatic in agreement. Standard deviation in responses was large for every question in this part in 2024 and the part about inbound and disposable texts (3c-3f) in 2023. The average responses therefore do not reflect a general consensus and may instead point to hesitation and confusion in making decisions based on end use.

4 Conclusion

Based on the survey results, it seems that the quality now achieved by MT is indeed considered relatively superior in terms of grammar. However, lexical and stylistic errors remain sites requiring post-editing by translation. As regards the distinction between light and full post-editing, responses do not emphatically support the idea that light PE may be sufficient for any translation that is to be published and also suggest that time and effort in light vs full PE continue to be significantly different. This seems to run counter to findings in other studies that find this distinction increasingly difficult to make. Finally, much reservation is expressed in doing away with PE. This is true in case of outbound texts, but also in case of inbound and disposable texts, albeit to a relatively lesser extent. The results suggest that PE is still considered essential for MT and that there remains a distinction between the extent of PE in spite of the progress achieved by MT.

Given the small size of respondents, it must be acknowledged that more large-scale surveys and those including other language pairs as well as professional translators would be necessary to

corroborate these inferences. It is also important to acknowledge that the fact that the respondents are training to become professional translators may have contributed to a bias and the hesitation reported. Studies have reported negative pre-task perceptions of MT contributing to lower quality and productivity in output (Briva-Iglesias & O'Brien, 2024, p. 451; Sánchez Ramos, 2025). The lack of experience in working with frameworks that clearly define requirements of quality and efficiency may also result in a conservative approach towards MT. This has been observed in previous studies (Mellinger, 2017; Venkatesan, 2023) and may affect responses of students.

References

Allen, J. (2003). Post-Editing. In H. Somers (Ed.), *Computers and translation: A translator's guide* (pp. 297–317). Benjamins.

Briva-Iglesias, V., & O'Brien, S. (2024). Pre-task perceptions of MT influence quality and productivity: The importance of better translator-computer interactions and implications for training. In C. Scarton, C. Prescott, C. Bayliss, C. Oakley, J. Wright, S. Wrigley, X. Song, E. Gow-Smith, R. Bawden, V. M. Sánchez-Cartagena, P. Cadwell, E. Lapshinova-Koltunski, V. Cabarrão, K. Chatzitheodorou, M. Nurminen, D. Kanojia, & H. Moniz (Eds.), *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)* (pp. 444–454). European Association for Machine Translation (EAMT). https://aclanthology.org/2024.eamt-1.37/

Goto, I., & Tanaka, H. (2017). Detecting Untranslated Content for Neural Machine Translation. *Proceedings of the First Workshop on Neural Machine Translation*, 47–55. https://doi.org/10.18653/v1/W17-3206

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., … Zhou, M. (2018). *Achieving Human Parity on Automatic Chinese to English News Translation*. https://doi.org/10.48550/ARXIV.1803.05567

Hutchins, W. J., & Somers, H. L. (1997). *An introduction to machine translation* (2nd ed.). Academic Press.

ISO 18587:2017 Translation services—Post-editing of machine translation output—Requirements. (2017). International Organization for Standardization. https://www.iso.org/standard/62970.html

Jia, Y., & Sun, S. (2023). Man or machine? Comparing the difficulty of human translation versus neural machine translation post-editing. *Perspectives*, *31*(5), 950–968. https://doi.org/10.1080/0907676X.2022.2129028

Läubli, S., Sennrich, R., & Volk, M. (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. *arXiv:1808.07048 [Cs]*. http://arxiv.org/abs/1808.07048

Massardo, I., Meer, J. van der, O'Brian, S., Hollowood, F., Aranberri, N., & Drescher, K. (2016). *TAUS Post-Editing Guidelines*. TAUS. https://commission.europa.eu/document/download/b482a2c0-42df-4291-8bf8-923922ddc6e1_en?filename=emt_competence_fwk_2022_en.pdf

Mellinger, C. D. (2017). Translators and machine translation: Knowledge and skills gaps in translator pedagogy. *The Interpreter and Translator Trainer*, *11*(4), 280–293. https://doi.org/10.1080/1750399X.2017.1359760

Nitzke, J., Canfora, C., Hansen-Schirra, S., & Kapnas, D. (2024). Decisions in projects using machine translation and post-editing: An interview study. *The Journal of Specialised Translation*, *41*, 127–148. https://doi.org/10.26034/cm.jostrans.2024.4715

Rico Pérez, C. (2024). Re-thinking Machine Translation Post-Editing Guidelines. *The Journal of Specialised Translation*, *41*, 26–47. https://doi.org/10.26034/cm.jostrans.2024.4696

Sánchez Ramos, M. D. M. (2025). Machine translation post-editing through emotional narratives: A methodological approach. *Translation and Translanguaging in Multilingual Contexts*, *11*(1), 31–47. https://doi.org/10.1075/ttmc.00152.san

Venkatesan, H. (2022). The fourth dimension in translation: Time and disposability. *Perspectives*, *30*(4), 662–677. https://doi.org/10.1080/0907676X.2021.1939739

Venkatesan, H. (2023). Technology preparedness and translator training: Implications for curricula. *Babel Revue Internationale de La Traduction / International Journal of Translation*, *69*(5), 666–703. https://doi.org/10.1075/babel.00335.ven

Wang, X., Wang, T., Muñoz Martín, R., & Jia, Y. (2021). Investigating usability in postediting neural

machine translation: Evidence from translation trainees' self-perception and performance: *Across Languages and Cultures*, *22*(1), 100–123. https://doi.org/10.1556/084.2021.00006

Appendix A. Questionnaire

Post-Editing Survey (2024)

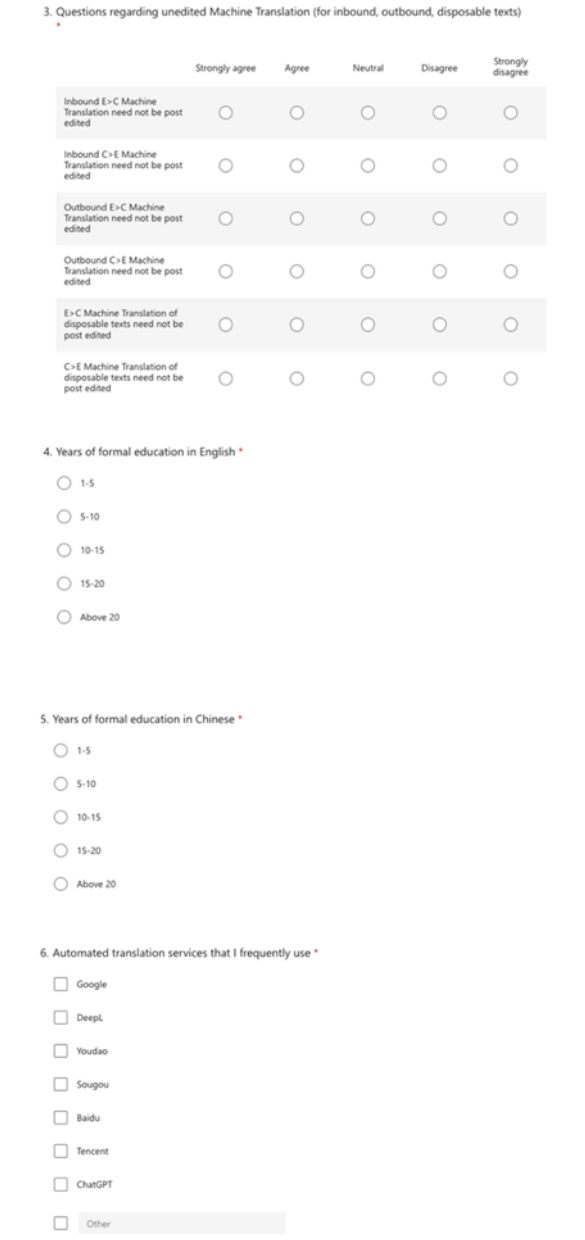
Please choose the best response for each of the following based on your experience of post-editing machine translation

1. Questions regarding efficiency and quality with post-editing machine translation. *

	Strongly Agree	Agree	Neutral	Disagree	Strongly disagree
Post-Editing E>C Machine Translation helps increase efficiency	○	○	○	○	○
Post-Editing E>C Machine Translation helps increase quality	○	○	○	○	○
Post-Editing C>E Machine Translation helps increase efficiency	○	○	○	○	○
Post-Editing C>E Machine Translation helps increase quality	○	○	○	○	○

2. Questions regarding types (grammar, lexis, style) and levels (light vs full) of post-editing required *

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
In E>C Machine Translation Post Editing is required to correct grammatical errors	○	○	○	○	○
In C>E Machine Translation Post Editing is required to correct grammatical errors	○	○	○	○	○
In E>C Machine Translation Post Editing is required to correct lexical errors	○	○	○	○	○
In C>E Machine Translation Post Editing is required to correct lexical errors	○	○	○	○	○
In E>C Machine Translation Post Editing is required to correct stylistic errors	○	○	○	○	○
In C>E Machine Translation Post Editing is required to correct stylistic errors	○	○	○	○	○
In E>C Machine Translation light post-editing is sufficient to make the text publishable	○	○	○	○	○
In C>E Machine Translation light post-editing is sufficient to make the text publishable	○	○	○	○	○
In E>C Machine Translation full post-editing is essential to make the text publishable	○	○	○	○	○
In C>E Machine Translation full post-editing is essential to make the text publishable	○	○	○	○	○
In E>C Machine Translation, time and effort taken to conduct light PE vs full PE is significantly different	○	○	○	○	○
In C>E Machine Translation, time and effort taken to conduct light PE vs full PE is significantly different	○	○	○	○	○



3. Questions regarding unedited Machine Translation (for inbound, outbound, disposable texts) *

	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Inbound E>C Machine Translation need not be post edited	○	○	○	○	○
Inbound C>E Machine Translation need not be post edited	○	○	○	○	○
Outbound E>C Machine Translation need not be post edited	○	○	○	○	○
Outbound C>E Machine Translation need not be post edited	○	○	○	○	○
E>C Machine Translation of disposable texts need not be post edited	○	○	○	○	○
C>E Machine Translation of disposable texts need not be post edited	○	○	○	○	○

4. Years of formal education in English *

- ○ 1-5
- ○ 5-10
- ○ 10-15
- ○ 15-20
- ○ Above 20

5. Years of formal education in Chinese *

- ○ 1-5
- ○ 5-10
- ○ 10-15
- ○ 15-20
- ○ Above 20

6. Automated translation services that I frequently use *

- ☐ Google
- ☐ DeepL
- ☐ Youdao
- ☐ Sougou
- ☐ Baidu
- ☐ Tencent
- ☐ ChatGPT
- ☐ Other

Appendix B. Standard Deviation in Responses (2023)

1a	0.587087048
1b	0.499432785
1c	0.499432785
1d	0.791107035
2a	0.791107035
2b	0.583211844
2c	0.583211844
2d	0.906014171
2e	0.906014171
2f	1.108613974
2g	1.108613974
2h	0.81092316
2i	0.81092316
2j	0.70950783
2k	0.70950783
2l	0.628138379
3a	0.628138379
3b	0.575383142
3c	0.575383142
3d	0.940400841
3e	0.940400841
3f	1.019092122

Appendix C. Standard Deviation in Responses (2024)

1a	0.476969601
1b	0.653834842
1c	0.591607978
1d	0.591607978
2a	0.653834842
2b	1.1
2c	0.589491306
2d	0.583095189
2e	0.792148976
2f	0.7
2g	1.235920709
2h	1.042832681
2i	0.734846923
2j	0.953939201
2k	0.90967027
2l	0.536190265
3a	1.24398553
3b	1.15758369
3c	1.314343943
3d	1.187434209
3e	1.280624847
3f	1.321930407

Is it AI, MT or PE that worry professionals: results from a Human-Centered AI survey

Miguel A. Jiménez-Crespo
Rutgers University
15 Seminary Pl., 5th fl. New Brunswick, NJ, 08901
jimenez.miguel@rutgers.edu

Stephanie A. Rodríguez
Rutgers University
15 Seminary Pl., 5th fl. New Brunswick, NJ, 08901
srodrig@newark.rutgers.edu

Abstract

Translation technologies have historically been developed without substantial input from professionals (e.g. O'Brien, 2012). Conversely, the emerging human-centered AI (HCAI) paradigm emphasizes the importance of including end-users in the "process of conceiving, designing, testing, deploying, and iterating" technologies (Vallor, 2024: 17). Therefore, early research engagement on the attitudes, needs and opinions of professionals on AI implementation is essential, as incorporating them at later stages "results in issues and missed opportunities, which may be expensive to recover from due to the cost, time, resources, and energy spent" (Winslow and Garibay, 2004: 123). To this end, this article presents a qualitative analysis of professional translators' attitudes towards AI in the future, centered around the role of MT and post-editing (PE). The discussion draws on data collected from open-ended questions included in a larger survey on control and autonomy from an HCAI perspective, which were thematically coded and qualitatively examined. The thematic analysis indicates that predominant concerns regarding the future of the AI-driven translation industry still revolves around longstanding issues in PE and MT literature, such as PE, translation quality, communicating and educating LSP, clients, users, and the broader public, as well as maintaining human control over the final product or creativity. This is explained to some extent to the relatively slow rate of integration of AI technologies into translation workflows to date (e.g. ELIS, 2025; Rivas Ginel et al., 2024; GALA, 2024, 2025; Jiménez-Crespo, 2024), or the fact the professional report using AI primarily for tasks related to translation, but not necessarily to PE the output of LLMs or NMT (Rivas Ginel and Moorkens, 2025).

1 Introduction

The launch of ChatGPT by the company OpenAI in November of 2022 started a revolution that was intended to transform a large number of fields (Raiaan et al., 2024). Large Language Models (LLMs) and different generative AI apps have been gradually implemented across professional fields, with translation and interpreting identified as an area of high exposure to negative impacts of AI (Eloundou et al., 20 23). In this context, concerns regarding the impact of AI have led to the emergence of the multidisciplinary field of Human-Centered AI (HCAI). This area of inquiry aims to position humans at the centered of technological developments (Ozmen Garibay et al., 2023), thereby ensuring that "their values and agency [are taken] into account" (Capel and Brereton, 2023: np). In countering the prevalent hype in the AI industry, HCAI represents "a paradigm shift, moving beyond the prevalent technology-centered approaches towards AI driven by human values" (Schmager et al., 2023: 7). A key issue addressed in this paper is that, even when AI and LLMs are supposed to revolutionize translation and interpreting practices, they are in fact not human-centered technologies (Vallor, 2024). Scholars have argued this because LLMs were developed without a clear focus on the needs, demands or preferences of existing end users. Instead, they emerged because evolving architectures and processing capabilities allowed companies, such as OpenAI, to successfully implement them (ibid). Nevertheless, they originally came without guardrails or clearly defined professional use-cases unsupervised use beyond the industry hype. This lack of human centeredness for professional tasks means that over the last two years, a large body of research

Proceedings of Machine Translation Summit XX Volume 1, pages 407–419, June 23-27, 2025

has been devoted to how, when or to what extent LLMs might be perceived as useful or and can be successfully integrated in professional tasks. In the language industry, Gen-AI and LLMs have been integrated (GALA, 2024, 2025; ELIS, 2025), often through trial and error and careful testing, in a wide range of tasks that include machine translation (MT), MT evaluation or Automatic post editing (APE). Both industry (GALA, 2024, 2025) and scholarly publications (Rivas Ginel and Moorkens, 2024) include a wide range of tasks in addition to translation. For example, recent studies have shown that professionals primarily use LLMs for tasks such as generating inspiration, summarizing content, rephrasing texts, understanding technical expressions, or performing terminology-related tasks (Rivas Ginel and Moorkens, 2024: 269). Nevertheless, translation is not reported as the most frequent use.

In this context, this paper reports on a qualitative section of a wider survey (Jiménez-Crespo, 2024) on attitudes towards the future impact of AI in three key areas of Human-Centered AI approaches, control, autonomy, and automation (Shneiderman, 2020, 2022). The need for this type of research is evident, as a key principle of HCAI approaches emphasizes the active participation of end-users throughout "process of conceiving, designing, testing, deploying, and iterating" technologies (Vallor, 2024: 17). Kishimoto, et al. also stress the importance of "involv[ing] potential users from the early stages of product and service development" because having an "inclusive R&D process is imperative" (2024: 3). They need to be incorporated in the early stages of AI development and deployment because incorporating them at later stages "results in issues and missed opportunities, which may be expensive to recover from due to the cost, time, resources, and energy spent" (Winslow and Garibay, 2024: 123). As AI technologies continue to advance, the understanding of user opinions and attitudes are critical for their successful adoption into the translation workflows. Such understanding helps to mitigate the risk of these technologies being perceived negatively, as imposed or restrictive by end-users (Ruokonen and Koskinen, 2017). These negative perceptions often lead to challenges with technology adoption and reduced job satisfaction (Sakamoto et al., 2024; Christensen et al., 2024).

The qualitative data analyzed for this paper focuses specifically on discourses by professionals surrounding machine translation post-editing (MTPE) on open ended questions related to future challenges posed by AI, as well as how automation might impact the techno sociological work conditions of translators in the USA. Published quantitative results (Jiménez-Crespo, 2024)[1] from the same survey study showed high self-reported levels of "perceived control" and "autonomy" over translation technologies, and subjects reported medium levels of forced technology use. Future perceived control in an AI era declined, but this perceived loss of control in the AI era was attributed to human agents in the process rather than AI apps or algorithms (big tech, developers, Language Service Provicers (LSPs), project managers, clients, etc.).

2 Methodology

This mix methods study involved a self-administered online Qualtrics survey available to professional US-based translators. The study obtained ethical clearance by Rutgers University ethical board and was piloted and revised. The survey was made available until June 15th, 2024 and 50 participants completed the survey. Participants were recruited online via e-mailings through all major professional associations in the US (e.g., ATA Language Technology Division, ATA Spanish Division, North-Eastern chapters of the American Translators Association) and social networks (e.g. LinkedIn). The only requirement to participate was to be a full-time translator in the USA with more than 2 years of experience. To encourage participation, a snowball sampling method was used (Goodman, 1961). Qualitative data in this paper were analyzed through thematic content analysis (Braun and Clark, 2006), utilizing a coding scheme that was developed inductively from emerging patterns in the data, then iterated, and finally used to categorize all responses. The bottom-up inductive analysis resulted on a coding scheme based on patterns in existing responses across the dataset. This initial set was used then by an additional researcher, and the coders then met to discuss any differences and to refine the scheme. Using this approach, the proportion of responses within each group that corresponded to a specific code were calculated, representing a theme identified in the dataset.

The survey had a final section with open-ended questions related to the future of AI-driven translation technology integrations in the HCAI

era. This section included five questions that provided the data analyzed in the present paper:

Question 23a: Human-Centered AI and the future Human-Centered AI involves a high degree of automation with humans firmly in control of the overall process. Imagine that in the near future you will work in a translation platform or translation management system powered by AI integrations. […] do you think you will have control over the integrations of AI in the translation process? Please explain
Question 25: Which part or subcomponents of the translation process do you think you might lose control over as AI becomes increasingly integrated into the workflow?
Question 26: If you had to provide input to design an AI technology tool to augment your capacities to translate better, more efficiently, or faster, how would you describe it?
Question 27: Human-Centered AI involves a high degree of autonomy of the human agent(s). If you would develop AI applications for translation, what would "autonomy" mean for you?
Question 28: In your opinion, what are the main challenges translators might face in the age of automation and AI?

All questions were optional, and participants could skip or not answer specific questions to avoid "survey fatigue" (Davis, 2019). The following responses were recorded for each question: Q23a= 25, Q25= 43, Q26= 35, Q27= 38, Q28= 41. The total responses recorded for open-ended questions dealing with an AI driven future were 182. The focus of the present analysis is on those themes and subthemes related to MTPE and MT, as well as conditions and issues related to these practices.

3 Results

3.1 Themes and subthemes

As previously mentioned, all responses to the open-ended questions were coded by the author and an additional researcher. The analysis of the dataset the author using thematic content analysis (Braun and Clarke, 2006). This resulted in 19 codes for themes and subthemes in those five questions related to the AI-driven future. The themes and subthemes are listed here in order of frequency.

- **PE**: References to post editing, either from NMT systems or LLMs.
- **Quality**: Issues related to translation quality of the final products or its implications.
- **Communication_edu_others**: Any issue related to how translators communicate or discuss the implications of using AI, NMT or other technologies with clients, LSPs, users or society at large. It includes issues related to perception of translators and translation in society, as well as the impact on their loss of professional recognition or status.
- **Replacement**: Any issue related to the potential replacement of translators by any type of technology.
- **Control_final**: Subtheme within the control theme related to the human control over the final product.
- **Tech_on_off**: Any reference to the ability of translators to activate or deactivate any type of technology for projects or at any point throughout the translation process.
- **Rates_competition**: Subtheme within the theme "Job Conditions", referring to the impact on translation rates or competition among translators that leads to reduced rates.
- **Creativity**: Reference to translation creativity.
- **Terminology**: Any reference to issues related to terminology during the translation process.
- **Transfer**: This is a subtheme related to PE, in which translators reference the ability to "transfer" the content or to produce the initial draft themselves, rather than being offered translation suggestions.
- **Adaptive_interactive**: References to adaptive or interactive technologies, both NMT or LLMs.
- **AI_companies**: References to AI or technology companies, typically relating to those in control of processes, development, and integrations.
- **Job_conditions**: References to job conditions of translators.
- **TM_improv_replacement**: References to TM either to improvements or to losing TM technologies due to AI.
- **Unsure**: Direct reference about respondents being unsure or unable to respond to a question that often appears in general survey studies on AI (e.g. Bingley et al., 2023).
- **Override_locked_segments**: This is a subtheme within the "PE" theme where translators discuss that they do not like locked

segments or the inability to override suggestions by NMT, TM, or AI.

- **Speed**: References to gains in translation speed using technology.
- **Human_superiority**: Direct reference to human superiority to machines on translation tasks.
- **Collaboration_with_devs**: References to the desire by translators to collaborate with developers of technologies to directly improve them.

Some other themes and subthemes that frequently appear in both TS and HCAI literature were less present in these responses, such as data biases (N=2), ethics (N=2), usability (N=3) or privacy (N=3).

3.2 Main themes: a summary

Table 1 shows the most frequent themes and subthemes for all questions, and here PE is not the most frequent theme in any of them. The second column, the summary, includes the aggregation of all values from all questions (R= 182). It includes the most frequent themes in all answers related to the future of the profession in the AI era. PE appears as the main theme overall for all, followed by quality, communication and education of other parties (clients, LSPs, users, developers, society), control over the final product and the ability to turn on and off technologies or to decide when to use them.

The rest of the columns show the most frequent theme in each question; For example, the main theme in Q26 (input to developers) is Adaptive_ interactive_tech. This theme does not refer exclusively to adaptive or interactive NMT technologies (Daems and Macken, 2019), as it also includes any type of "adaptation" including the ability of AI implementations to adapt to different contexts, genres, registers, or even dialectal variation. Thus, it includes adaptation both to user preferences and to text-specific issues. Other themes that frequently appear include communication and education with clients, end-users, LSPs or society at large for Q28 related to future AI challenges. Q25 related to what might be lost in the AI era showed that the preservation of human creativity was the most important theme, while for Q27 related to what "autonomy" means in the AI-driven future the main theme was human control over the final product. Finally, in Q23a that requested additional information on whether translators will retain control in the AI era, the theme AI companies was the main theme. This last issue aligns with findings from previous research (Jiménez-Crespo, 2024) that translators place the blame on human agents for losing control and autonomy regarding technological decisions rather than AI technologies themselves, such as AI companies, AI, and translation tech developers, LSPs, translation managers or workflow designers.

3.3 What is lost with AI? From "transfer" to PE

In question Q25, related to what might be lost with future AI integrations, a subtheme within the PE theme was identified that was labeled as "transfer". The three most frequently identified themes and subthemes in participants' responses to what will be lost with AI were "creativity", "transfer", and "PE". In the iterative analysis to identify the themes and subthemes, it was decided that "transfer" represented a subtheme within the "PE" theme because both "PE" and "transfer" represent two sides of the same coin. Depending on the question, the perceived loss of the ability to "transfer" the initial translation or whether translators will lose the ability to produce the translation from scratch represents the same theme from a different perspective related to how translators cognitively process translations. This shift from traditional translation from scratch to PE is thus frequently described as a "loss." (e.g. Pielmeier and O'Mara, 2020; Girletti, 2024). Notably, translation scholars have always emphasized that translation involves a "transfer" stage. From a theoretical perspective, Gideon Toury (1995) proposed three postulates of translation or what "translation" is: the (1) source and (2) target text postulates, as well as the (3) 'transfer" one, underscoring that translation proper requires a "transfer" stage. Studies have delved into whether automatic transfer, followed or not by PE, can be considered as "translation". Similarly, resistance by professionals to the practice of PE is based on the premise that automatic transfer is not conceptualized a type of "translation". Thus, in this Q25, the subtheme "transfer" was identified in 21.73% answers, while the wider theme "PE" represented 15.94% of the tagged themes identified. Across all survey questions analyzed in this paper, these themes represented 10.68% and

Most frequent themes	Perceptions towards AI-driven future Summary	Q. 28. AI Challenges (R=41)	Q25.What parts of the process will be lost (R=43)	Q26. Input to design augmented tech (R=35)	Q27. Autonomy in an HCAI future (R=38)	Q23a. Future control in HCAI age (R=25)
N1	1. PE	1.Comm_Edu	1.Creativity	1. Adapt_ interact_tech	1. Control_ Final	1. AI_ companies
N2	2. Quality	2. Quality	2.Transfer	2.Configuratio n	2. Tech_on_ off	2. Term
N3	3. Comm_Edu	3. Replacement	3. PE	3. Unsure	3. PE	3. Unsure
N4	4. Control_Final	4. Rates_ Competition	4. Quality	4. Usability	4. Privacy	4.Diff_ workflow integration_ process
N5	5. Tech_on_off	5. PE	5. Term	5. Speed	5. Configure	

Table 1: Summary of most frequent themes in each open-ended question related of the AI driven future. R indicates number of responses for each question, while N indicates the order of frequency for each theme (N1-N5).

6.47%, respectively. Notably, one key finding is that depending on how questions are phrased, responses refer to PE or transfer as the terminology of choice, even when though both terms might describe to the same notion of professionals not translating without prepopulated translation candidates. Participant P34 directly addressed this issue when responding to a question about what professionals might lose in the age of AI:

- The power to negotiate fare rates, the ability to translate from scratch if all the agencies are asking is postediting, quality of the final result (P34) [emphasis own]

This response also addressed other key themes, such as "quality" and "rates". This sense of loss in translation, conceptualized as the inability to craft the initial round of translation, is described by respondents as losing "the actual conversion of one language to another" (P15), the "translation step" (P45) or "the act of translating. I feel humans will become proofreaders" (P14). This is often conceptualized negatively, such as the following response indicating not only that LSPs will require the use of technology, but "even worse", LSPs will present to translators pre-processed files with AI:

- I'm expecting it will be integrated into tools that LSCs will try to require use of. *Even worse would be receiving pre-processed files (segments pre-populated and sometimes locked for editing) where the pre-processing is automatically generated from AI* (rather than TMs) (P20) [emphasis own].

The last words of this response related to phasing out TM technologies is addressed in another subtheme in the analysis. This is perceived as a potential loss in the AI age, as TM technologies are perceived as reflecting human contributions. This subtheme "Losing_TM" within the overall "TM" theme represents 2.8% of overall themes (N=3). Participants described the "transfer" theme in various ways, but it is most often identified with the initial or first phase of translation:

- The initial production of a draft (P16)
- Initial translation and possibly final product (P20)
- The first translation step (P45)
- The initial round of translation, also the ability to override a machine's -approval/acceptance of a translated segment/term/usage/grammar etc. (P41)

In these formulations it can be perceived that respondents often indicate that "translation" will be lost, signaling that PE might not be translation at all. This perception of losing the ability to "transfer" is often related to the second most frequent theme identified in this question, losing "creativity" or the creative potential of the translator:

> -Translating! *AI is not creative, and I work in creative fields of translation.* I don't want to see AI
> suggestions, because they will block my own creativity (studies have shown this to be true). So I am not interested in integrating AI into my workflow. I intend to produce "hand-crafted" translations as long as I can, and I think I work in fields where this approach is valued (P43) [emphasis own].

The fact that PE leads to the loss of creativity has been identified in previous PE studies (e.g. Álvarez-Vidal, Oliver and Badia, 2020), and it is recently one of the most popular research trends within a multidisciplinary area that includes translation studies, literary studies, and computational linguistics (e.g., Guerberof-Arenas and Toral 2020, 2022; Toral and Guerberof-Arenas, 2024; Winters and Kenny, 2023; Kenny and Winters, 2024; Resende and Hadley, 2024).

3.4 Control or autonomy

Control and autonomy are two cornerstones of HCAI approaches (Shneiderman, 2020, 2022). Q27 directly addressed what autonomy might mean in the AI-driven future. The analysis reveals that autonomy is conceptualized in terms of whether translators retain full control of the range of technologies they use (or not), and whether these technologies are imposed by third parties, such as LSPs or AI companies. The role of LSPs and key stakeholders in determining the adoption and implementation PE practices and how it has been previously studied, for example, Nitzke et al. (2024), detail the factors that influence workflow decisions regarding MTPE that are subsequently imposed on participating professionals. In literary translation, Way et al. (2024: 97) stress the importance for practitioners to retain human "control over their preferred translation workflow" and whether to include MT. In this regard, one key finding in this study is that translators perceive their autonomy in the translation process often in terms of whether they can reject any work that involves the imposition of any tool (select_reject_work):

- I can turn away work that requires me to use tools I don't want to work with (P1)
- I don't work for clients who control my technology (P18)

This is often conceptualized in terms of the ability to make their own decisions rather than having choices imposed upon them:

- Ability to decide which ones are better and when, and not to depend on clients or others to impose (P29)

The reasons why freelancers often conceptualized autonomy as the ability to reject or select work assignments are related to not having access to certain technologies if they are not provided by the LSP, or even that from a usability standpoint they do not feel comfortable using:

- I do not accept assignments that require use of technology I don't have access to or am not comfortable using (P44)

Participants, thus, showcase what has been shown in the study by Nitzke et al. (2024) with stakeholders in making MTPE decisions that "working conditions and prospects for highly qualified and technology-savvy translators in the high-end segment are good despite, claims to the contrary (2024: 143).

Control and autonomy extend to the most frequent theme in Q27, the ability to retain control over all features of the final product (control_final) as observed in previous studies (Rossi and Chevrot, 2019; Girletti, 2024). Regardless of whether PE is used, in combination with AI solutions or independently, respondents indicated that their autonomy would only be considered respected in the future if they retain their agency and decision-making ability in all aspects of the final product.

- Make the final decision for all the steps of the translation (P29)

For example, these respondents indicate that they welcome AI suggestions in the translation process,

but they would like to have the final say in the translation.

- The AI is really just making suggestions; "autonomy" is me creating the translation (P44).
- Be able to create the translation from scratch, with the AI assisting me with research in context (P34)

In some instances, respondents continue to welcome automation and AI assistance, yet they express their resistance to segments that are machine evaluated and automatically approved:

- That nothing is translated without the user clicking a check box to indicate the translation is human approved (P31)

This ties with one of the subthemes identified within the PE theme, the override_locked_segments:

- As the translator, to be able to change anything you didn't think was correct (P28)
- […] the ability to override a machine's approval/acceptance of a translated segment/term/usage/grammar etc. (P41)

In addition, one respondent (P42) indicates that nothing is automatic given that humans set the parameters for automation:

- Nothing is automatic, all autonomy is first decided by a human (P42)

This response is related to the previously mentioned issue that respondents always blame other human agents for their perceived lack of autonomy and control. In this regard, another of the key themes of this area is the ability to control how and when technology is implemented for specific projects, translations or throughout the day (tech_on_off). This ability to integrate different technologies depending on human cognitive or processing demands, based on user preferences or psychophysiological status, emerges as a key theme in this study. For example, respondents indicate:

- The freedom to select which components I incorporate into my workflow, and the extent to which such components are incorporated in any given project (46)
- Autonomy in my view means: 1) Ability to activate/deactivate functionalities […] (P9)

It also implies that during any specific passage, moment or part of a project assistance could be turned on or off:

- Autonomy for me would mean that with a click of a button I could turn AI intervention on or off (P43)

Here, a key issue in AI augmentation approaches is the need for those integrating technologies into current or future workflows to establish "which tasks to automate, which tasks to augment, and which tasks to leave to humans" (Sadiku and Musa, 2021: 191). In practice, decisions related to the levels of automation are often made by LSPs and/or translation managers. However, professionals prefer for the locus of control to reside in themselves, being able to decide when to PE, when to translate from scratch, or when and how to integrate LLM suggestions. This, as Ruffo and Macken indicate, might be more important than any time or efficiency gains for literary translators (Ruffo and Macken, 2024: 241).

3.5 Adaptive or Interactive MT and AI technologies

Adaptive or interactive MT has been one of the key technological developments prior to the emergence of LLMs (e.g. Daems and Macken, 2019; Daems, 2024; Briva Iglesias and O'Brien, 2023). In Question 26, participants were asked to identify input features they would like to provide to AI developers to design technologies that would augment their capacities to translate better or more efficiently. Interestingly, adaptive or interactive MT capabilities emerged as the most frequently mentioned theme among respondents. For instance, one respondent indicated in a brief response "Adaptive AI" (P45) or "Off-line and adaptable translation" (P2). This is also expressed in the following fashion:

- MT that adapted based on the way I post-edited it in a previous segment of current job or of previous translation job (P33)

- I can't envision anything outside better translation memories. I would like AI to remember how I translated individual words or phrases (P31)

Again, interaction closely relates to the previous subtheme related to locked segments or the ability to override AI decisions:

- It should be interactive rather than "over the fence" or post-editing. Offer suggestions rather than assume you will accept it 100% (P20)
- …Offers flexibility: Functions can be activated/deactivated at will… 3) Can interact with external sources:…. Gives the translator freedom to edit the target language as he/she wishes… (P9)

The adaptive capabilities of the MT or AI system should extend not just to the interaction and adaptation to the user, but also to the type of text and genre. Thus, the ability to make context-specific suggestions or choices emerges as a subtheme within this adaptive/interactive theme.:

- I would like to see AI tools that can make choices based on context - time, place, type of document, language register, etc. (P1)
- To grasp more context (P14)

This ability to adapt to context also extends to a key issue in languages with multiple dialectal varieties, specifically the ability to help in dealing and adapting to specific language varieties (Jiménez-Crespo and Rodríguez, 2024).

- One of the things I struggle with, is that Spanish is spoken differently across the world. So AI translates for one word that may not be used in some of these countries. An optional translation tool would be great. (P47).

3.6 Other themes related to PE and MTPE: from lower rates to replacement or collaboration to develop tools

Several additional themes relate to MTPE and appear in published survey-based literature, such as concerns regarding reduced rates through a combination of PE and AI app integration (e.g. Laübli and Orrego-Carmona, 2017; Caldwell, O'Brien and Teixeira, 2018; Alvarez-Vidal, Oliver and Bandia, 2020; ELIS 2025). As indicated in the 2024 ELIS report (2024), professionals normally conflate both MT and AI to blame for lower rates as "AI and MT are considered to be equivalent in the sense that both reduce the appreciation and therefore also the financial compensation, for human language work" (ELIS 2024: 40). This is perceived in the analyzed data, with some participants explicitly linking the perceived future threat of AI integration to MTPE, particularly due to its potential effect of lowering translation rates:

- Clients might approach translators with machine post-editing assignments rather than translation jobs to save money (P7)

Other participants report this fear of lower rates with the fear of replacement in some tasks:

- Economic challenges: a tighter market for translators with lower rates. […] Now translators will be hired for less money to revise or check AI writing or translation (P43).

In some cases, this fear of lower rates is also connected to fear of replacement and the disappearance of professional translation work:

- I think the main challenge will be to have a job to do. If companies go totally for AI without human control or humans post editing the translations, then there will be no jobs (P47)

While others blame potential lower rates with the hype of the industry on the abilities of AI apps.

- Downward pressure on rates without commensurate gains in efficiency or reductions in actual labor expenditure due to overblown confidence in the capacity of AI. Indeed, a bad tool can often *reduce* efficiency or *increase* labor, if my experiences with MTPE are any indication. (P46)

Nevertheless, a recent study does not show a rate reduction in the AI age with 70.34% of respondents to recent survey indicating similar or increasing rates (Rivas Ginel et al., 2024), while other surveys have shown otherwise (e.g., ELIS 2025). In addition, as shown by other studies, the fear or

lower rates is related to competition by other translators that accept certain conditions that impact across the board:

- Lower and lower rates for translation (translators using AI accept lower rates and that lowers the rates across the board). (P31).

Nevertheless, even when the attitudes towards PE and automation in the data are mostly negative, some positive attitudes are still also found:

- Many translators also feel like automation and AI is here to steal their livelihood. I personally don't feel that way, as I understand automation can be good if we have a voice in how it's implemented. (P25)

In any case, this positive attitude is directly connected to the ability to control automation and how it is implemented. As indicated in a recent study on claims of AI augmentation in collaborative platforms by Jiménez-Crespo (2023), translators can only be augmented from a HCAI perspective if they locus of control resides in human participants, and they fully retain their agency.

A final theme of interest is the call from professionals to collaborate with developers, a key issue for technology to be human centered (Vallor, 2024; Schmager et al., 2023). This is something that has not happened historically (O'Brien, 2012), with developers of MT systems more interested in efficiency gains over creating human centered tools. This seems to be a trend that is starting, as reported in Rivas Ginel and Moorkens (2024) but it is still a desirable position for translators.

- I could be wrong, but I get the impression AI companies are not including translators in conversations related to design and functionality, but they only want translators to do language proofreading to help perfect AI's language output (P15)

Translators thus would like to be part of the development process beyond having their output used for training systems and extending it to user experience and user interfaces (Briva-Iglesias and O'Brien, 2024).

4 Limitations of the study

The study had certain limitations. First and foremost, the size of the sample. As previously mentioned, the call for participation was posted on the main professional forums in the USA and several chapters of the American Translators Association. It is possible that both the extensive nature of the survey and the theoretical approach that focused on certain aspects related to HCAI, control and autonomy might have discouraged potential participants or prevented them from completing the survey. In addition, no direct compensation was offered for participation. Second, it is possible the "survey fatigue" (Davis, 2019) might have influenced response rates, given the large numbers of national and international AI survey studies. Nevertheless, it can be argued that the study is representative of the targeted population to the extent that some results overlap with similar much larger surveys in Europe. For example, the ELIS (2024) and the Rivas Ginel et al. (2024) survey identified a 37-40% use of AI and LLMs by professionals in mid 2024, the same as the present study (Jiménez-Crespo 2024). While the relatively low response rate might be due to a methodological issue in the instrument design, the substantial data compiled provides a clear snapshot of the current attitudes towards AI. In terms of replicability, the complete survey is already accessible in an open science, freely accessible journal for replicability in other regions or settings (Jiménez-Crespo 2024).

5 Conclusions

This qualitative study explored the attitudes towards AI, particularly focusing on HCAI issues, such as control and autonomy, in the context of MTPE and the MT capabilities of recent AI driven LLM models. The survey was responded by 50 US-based professionals and the present study focused on five open ended questions about the future impact of AI on their profession, future job conditions, autonomy or how they envision an "augmented" future.

Overall, the results show that the main future challenges and attitudes towards AI technologies in the AI era primarily center on PE, control over the initial transfer process from the source text, and translation quality. This is followed by themes, such as communicating and educating LSP, clients, users, and society at large, human control over the

final product and the ability to turn on and off technologies or decide when and how to use them. Other less frequent themes emerge as key areas of concern depending on the question posed to respondents, such as creativity, and the attitudes toward adaptive or interactive technologies. When asked about the main challenges posed to translators, issues such as human replacement or rates also appear as key themes. In summary, the thematic analysis of the dataset reveals that those current concerns regarding the AI-driven future still revolve around established issues in PE and MT literature.

Notably, several themes and subthemes that frequently appear in both TS and HCAI literature were less present in these responses towards the future of the profession in an HCAI era, such as data biases, ethics, or usability. Current attitudes toward an AI-augmented future remain predominantly characterized by established concerns, such as resistance to PE, questions about whether relinquishing the initial draft translation (the transfer stage) fundamentally alters the essence of "translation," and related implications for quality, compensation rates, and creative expression. In terms of what professionals' expressed demands, the main themes in the data revolve around developing adaptive or interactive MT and LLM technologies and the full ability to control the final product without impositions such as locked segments, terminology, or the ability to control or override any AI implementations. This study addresses calls for further research into translators' attitudes towards translation technology in MT and AI era (Sakamoto et al., 2024; Christensen, Bundgaard and Flanagan, 2024), and demonstrates the need of a human-centered approach to foster translators' well-being, satisfaction and rates of adoption of technologies.

References

Álvarez-Vidal, Sergi, Oliver, Antoni, and Toni Badia. 2020. Post-editing for Professional Translators: Cheer or Fear?. *Revista Tradumàtica. Tecnologías de la Traducción,* 18: 49-69. https://doi.org/10.5565/rev/tradumatica.275

Axente, M., Eds.; Chapman and Hall/CRC, 2024; pp. 13–20. Way, Andy, Rothwell, Andrew and Roy Youdale. 2023. Why Literary Translators should embrace Translation Technology. *Revista Tradumàtica. Tecnologies de la Traduccio,* 21:87–102

Bingley, William J. et al. 2023. Where is the human in human-centered AI? Insights from developer priorities and user experiences. *Computers in Human Behavior*, 141: 107617.

Braun, Virginia and Victoria Clarke. 2026. Using thematic analysis in psychology. *Qualitative Research in Psychology,* 3: 77–101.

Briva-Iglesias, Vicent, O'Brien, Sharon, and Cowan, Benjamin R. 2023. The impact of traditional and interactive post-editing on machine translation user experience, quality, and productivity. *Translation, Cognition and Behavior*, 6(1): 60-86.

Briva-Iglesias, Vicent and Sharon O'Brien. 2024. Pre-task perceptions of MT influence quality and productivity: the importance of better translator-computer interactions and implications for training. *The 25th Annual Conference of The European Association for Machine Translation*. Sheffield, University.

Christensen, Tina P., Bundgaard, Kristin and Marian Flanagan. 2024. Psychological consequences of the digital transformation of the translation industry: an exploratory study of technostress among Danish certified translators. *Tradumàtica tecnologies de la Traducció,* 22: 187-206

Cadwell, Patrick, O'Brien, Sharon, and Carlos S. C. Teixeira. 2018. Resistance and Accommodation: Factors for the (Non-) Adoption of Machine Translation among Professional Translators. *Perspectives: Studies in Translatology*. 26(3), 301-321.

Capel, Tara and Margot Brereton. 2023. What is human-centered about human-centered AI? A map of the research landscape. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 2023, 1–23.

Daems, Joke. 2024. Students' Attitudes Towards Interactive and Adaptive Translation Technology: Four years of Working with Lilt. *In New Advances in Translation Technology: Applications and Pedagogy* (pp. 239-261). Singapore: Springer Nature Singapore.

Daems, Joke and Lieve Macken. 2019. Interactive adaptive SMT versus interactive adaptive NMT: a user experience evaluation. *Machine Translation*, 33(1): 117-134.

Davies, Jack. 2019. Think you're sending too many surveys? How to avoid survey fatigue. *Qualtrics Blog*. https://www.qualtrics.com/blog/avoiding-survey-fatigue/

ELIS 2024. ELIS Research. 2024. *European Language Industry Survey 2024.* Brussels: European Union of Associations of Translation Companies. https://elis-survey.org/wp-content/uploads/2024/03/ELIS-2024-Report.pdf

ELIS Research. 2025. *European Language Industry Survey 2025.* Brussels: European Union of Associations of Translation Companies. http://elis-survey.org/wp-content/uploads/2025/03/ELIS-2025_Report.pdf

Eloundou, Tina, Manning, Sam, Mishkin, Pamela, and Daniel Rock. 2023. GPTs are GPTs: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv*:2303.10130.

GALA. 2024. AI and Automation Barometer Report 2024. *GALA, Globalization and Localization Association*. https://www.gala-global.org/knowledge-center/professional-development/articles/ai-automation-barometer-report

GALA. 2025. Technology, Automation and AI. *GALA, Globalization and Localization Association.* https://www.gala-global.org/knowledge-center/professional-development/articles/gala-business-barometer-technology-ai-and

Ozmen Garibay, Ozlem, et al. (2023). Six human-centered artificial intelligence grand challenges. *International Journal of Human–Computer Interaction*, 39 (3): 391-437.

Girletti, Silvia. 2024. Beyond the assembly line: exploring salaried linguists' satisfaction with translation, revision and PE tasks. *Revista tradumàtica: traducció i tecnologies de la informació i la comunicació*, 22: 207-237.2024.

Goodman, Leo A. 1961. Snowball sampling. *The annals of mathematical statistics* 32: 148-170.

Guerberof-Arenas, Ana and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience". *Translation Spaces*, 9(2): 255-282. https://arxiv.org/abs/2101.06125

Guerberof-Arenas, Ana and Antonio Toral. 2022. Creativity in Translation: Machine Translation as a Constraint for Literary Texts. *Translation Spaces*, 11 (2): 184–212. https://doi.org/10.1075/TS.21025.GUE.

Jiménez-Crespo, Miguel A. 2023. Augmentation in Translation Crowdsourcing: Are Collaborative Translators' Minds Truly 'Augmented'? *Translation, Cognition and Behavior*. http:// 10.1075/tcb.00079.jim

Jiménez-Crespo, Miguel A. 2024. Exploring professional translators' attitudes towards control and autonomy in the Human-Centered AI era: quantitative results from a survey study. *Tradumatica: Translation Technologies. Special issue on Study on Human-Computer Interaction in Translation and Interpreting: Software and Applications* 22: 276-301.

Jiménez-Crespo, Miguel A., and Stephanie Rodríguez. 2024. Spanish Translation and the Role of Machine Translation and AI Technologies for Public Communication in the United States. *Estudios del Observatorio*, (93): 73-87.

Kenny, Dorothy and Marian Winters. 2024. Customization, personalization and style in literary machine translation. *Translation, Interpreting and Technological Change: Innovations in Research, Practice and Training*, 59.

Kishimoto, Atsuo, et al. 2024. Introduction. In Catherina Regis, Denis, Jean Louis, Axente, Maria L. and Atsuo Kishimoto, Eds. *Human-Centered AI: A Multidisciplinary Perspective for Policy-Makers, Auditors, and Users* (pp 1-12). CRC Press, 2024.

Läubli, Samuel, and David Orrego-Carmona. 2017. When Google Translate Is Better than Some Human Colleagues, Those People Are No Longer Colleagues. *Proceedings of the 39th Conference Translation and the Computer*, pp. 59-69

Nitzke, Jean et al. 2024. Decisions in projects using machine translation and post-editing: An interview study. *The Journal of Specialised Translation*, 41: 127–148.

https://doi.org/10.26034/cm.jostrans.2024.4715

O'Brien, Sharon. 2012. Translation as human–computer interaction. *Translation Spaces,* 1(1): 101-122.

Pielmeier, Helene and Paul O'Mara. 2020. *The State of the Linguist Supply Chain: Translators and Interpreters in 2020*. Common Sense Advisory Research. <https://cdn2.hubspot.net/hubfs/4041721/Newsletter/The%20State%20of%20the%20Linguist%20Supply%20Chain%202020.pdf>.

Raiaan, Mohaimenul A. K. et al. 2024. A review on large Language Models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*.

Resende, Natalia and James Hadley. 2024. The translator's canvas: Using LLMs to enhance poetry translation. *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas* (Volume 1: Research Track).

Rivas Ginel, Maria I., and Sader Feghali, Lina and Francesca Accogli. 2024. Exploring Translators' Perceptions of AI. *ELC Survey*. http://dx.doi.org/10.13140/RG.2.2.23582.75842

Rivas Ginel, Maria A., and Joss Moorkens. 2024. A year of ChatGPT: translators' attitudes and degree of adoption. *Tradumática*. *Tecnologies de la Traducció*, 22: 258-275. https://doi.org/10.5565/rev/tradumatica.369

Rossi, Caroline and Jean-Pierre Chevrot, 2019. Uses and perceptions of machine translation at the European Commission. *The Journal of specialised translation (JoSTrans),* 31: 177-200.

Ruffo, Paola, Dames, Joke and Lieve Macken 2024. Measured and perceived effort: assessing three literary translation workflows. *Revista Tradumàtica. Tecnologies de la Traducció*, 22: 238-257. https://doi.org/10.5565/rev/tradumatica.378

Ruokonen, Minna and Kaisa Koskinen. 2017. Dancing with technology: Translators' narratives on the dance of human and machinic agency in translation work. *The Translator*, 23: 310–323.

Sadiku, Mathew and Sarham Musa. 2021. A primer on multiple intelligences. Springer.

Sakamoto, Akiko; Van Laar, Darren; Moorkens, Joss; do Carmo, Félix. 2024. Measuring translators' quality of working life and their career motivation: conceptual and methodological aspects. *Translation Spaces*, 13(1): 54-77

Schmager, Stefan et al. 2023. Defining Human-Centered AI: A Comprehensive Review of HCAI Literature. In *Proceedings of the Mediterranean Conference on Information Systems*.

Shneiderman, Ben. 2020. Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction* 12(3): 109–124. https://doi.org/10.17705/1thci.00131.

Shneiderman, Ben. 2022. *Human-centered AI.* Oxford: Oxford University Press.

Teixeira, Carlos. 2014. Perceived vs. measured performance in the post-editing of suggestions from machine translation and translation memories. Proceedings of the *Third Workshop on Post-editing Techniques and Practices (WPTP-3): The 11th Conference of the Association for Machine Translation in the Americas: Vancouver, BC Canada*. AMTA, pp. 45-59.

Toral, A. and Guerberof-Arenas, A., 2024. "To Be or Not to Be: A Translation Reception Study of a Literary Text Translated into Dutch and Catalan Using Machine Translation." *Target*. https://doi.org /10.1075/target.22134.gue

Toury, Gideon. Descriptive Translation Studies. Amsterdam: Joyn Benjamins.

Vallor, Shannon. 2024. Defining human-centered AI: An interview with Shannon Vallor. In Catherina Regis, Denis, Jean Louis, Axente, Maria L. and Atsuo Kishimoto, Eds. *Human-Centered AI: A Multidisciplinary Perspective for Policy-Makers, Auditors, and Users;* CRC Press, 2024.

Way, Andy, Roy Youdale, and Andrew Rothwell. Why more literary translators should embrace translation technology. Tradumatica, 21: 87-102. https://doi.org/10.5565/rev/tradumatica.344

Winters, Marin and Dorothy Kenny. 2023. Mark my keywords: a translator-specific exploration of style in literary machine translation. *In Computer-Assisted Literary Translation* (pp. 69-88). Routledge.

Winslow, Ben, and Ozlem Garibay O. 2024. Human-Centered AI. In *Human-Computer Interaction in Intelligent Environments* (pp. 108-140). CRC Press.

Prompt engineering in translation: How do student translators leverage GenAI tools for translation tasks

Jia Zhang[a] **Xiaoyu Zhao**[b] **Stephen Doherty**[a]
a. School of Humanities and Languages, The University of New South Wales
b. Monash Suzhou Research Institute, Monash University
jia.zhang2@unsw.edu.au xiaoyu.zhao@monash.edu s.doherty@unsw.edu.au

Abstract

GenAI, though not developed specifically for translation, has shown the potential to produce translations as good as, if not better than, contemporary neural machine translation systems. In the context of tertiary-level translator education, the integration of GenAI has renewed debate in curricula and pedagogy. Despite divergent opinions among educators, it is evident that translation students, like many other students, are using GenAI tools to facilitate translation tasks as they use MT tools. We thus argue for the benefits of guiding students in using GenAI in an informed, critical, and ethical manner. To provide insights for tailored curriculum and pedagogy, it is insightful to investigate *what* students use GenAI for and *how* they use it. This study is among the first to investigate translation students' prompting behaviours. For thematic and discourse analysis, we collected prompts in GenAI tools generated by a representative sample of postgraduate student participants for eight months. The findings revealed that students had indeed used GenAI in various translation tasks, but their prompting behaviours were intuitive and uninformed. Our findings suggest an urgent need for translation educators to consider students' agency and critical engagement with GenAI tools.

1 Generative AI and Translation

AI has gradually permeated our life and work over the past two years. In particular, the launch of ChatGPT in 2022 captured significant attention across various sectors with its unprecedented ability to generate contextually relevant responses based on pattern recognition. Since then, ChatGPT and other Generative AI (GenAI) tools have experienced rapid development and continued to attract public attention. GenAI tools have now been embedded in our smartphones and laptops with great utility. Despite their limitations, GenAI tools are also said to have significantly transformed our work and the industries at large by improving automation, efficiency and productivity (McKinsey & Company, 2023).

In the industry and discipline of translation and interpreting, GenAI has also been experimented with and adopted by language service providers and professional translators. Though not specifically developed for translation, GenAI has been applied to converting texts from one language to another, given the training data and neural network architecture similarities between GenAI and Neural Machine Translation (NMT). Both GenAI and NMT rely on natural language processing and transformer-based models. GenAI has shown the potential to generate translations of quality equal to, if not superior to, that of contemporary NMT (Lee, 2023). Thus, we argue that GenAI tools can be considered a broader form of MT and language tools.

However, automatically translating from one language into another is merely one of GenAI's many functions. Beyond automatic translation, GenAI has been instrumental in facilitating the entire translation process, from background information searching and translation strategy analysis to proofreading and editing. Consequently, there is a growing trend among professionals to integrate GenAI into translation workflows, exploring innovative ways to enhance translation productivity and quality.

Indeed, the role of GenAI tools, especially ChatGPT, in empowering human translators has been discussed and researched in the last two years. Studies have shown that GenAI offers advantages over human translators in terms of efficiency in

Proceedings of Machine Translation Summit XX Volume 1, pages 420–431, June 23-27, 2025

processing lengthy text, accuracy in terminology translation, and consistency in style (e.g., Fu & Liu, 2024; Mohammed et al., 2024; Tekwa, 2024). When collaborating with human translators, GenAI models have outperformed contemporary NMT models (e.g., Google Translate) in enhancing translation quality by integrating pre-editing analyses and interactive inputs (e.g., Wu et al., 2023). While GenAI has proven effective in assisting translation practices, it also exhibits significant shortcomings, such as accuracy issues (e.g., mistranslations from limited contextual understanding) (Mohsen, 2024), creativity constraints (e.g., failure to produce nuanced and culturally resonant translations) (Katan, 2022), and ethical concerns (e.g., perpetuation of biases in training data) (Jiménez-Crespo, 2024). Addressing these challenges requires human discretion in critically evaluating AI outputs (Katan, 2022).

2 GenAI and translator training

In the context of tertiary-level translator education, the integration of GenAI has renewed previous debates on the benefits and challenges of integrating translation technologies, particularly machine translation, into our curricula (e.g., see Doherty, 2016; Doherty & Moorkens, 2013; Kenny & Doherty, 2014). On the one hand, the integration of GenAI tools into translator education has been advocated, given its benefits (Zhang, 2025), which have been explored in previous studies, including improving bilingual and extra-linguistic competencies and enhancing translation efficiency (e.g., Li & Tian, 2024). On the other hand, the inappropriate integration of GenAI into translator education could adversely affect the development of students' translation competence. Given the current limitations of GenAI-generated translations, students must acquire critical skills to evaluate and refine these outputs. However, translation students' overreliance on GenAI during the learning process may raise concerns about the non-critical evaluation and use of its outputs (Li & Tian, 2024). However, translation students' overreliance on GenAI during the learning process may raise concerns about the non-critical evaluation and use of its outputs (Li & Tian, 2024).

Regardless of the debate concerning the integration of GenAI in translator training, the lack of empirical studies means that most discussions and decisions about GenAI in translator training are experiential and intuitive. So far, the integration of GenAI in translator training has been extensively discussed, mainly in theoretical literature. Scholars tend to focus on how technology impacts translator training and what transformation is needed for translation programs (e.g., Li et al., 2023; Zhao et al., 2024). There is further discussion on how GenAI can be leveraged to teach translation and technology. However, relevant empirical studies are scarce, with only a handful of survey-based studies investigating students' and teachers' perceptions of AI in translation (e.g., Łukasik, 2024; Sahari et al., 2023). Evidence regarding students' interaction with GenAI, such as their prompting strategies, or the effect of teaching with GenAI, has yet to be found.

Indeed, these issues in the debate regarding integrating AI in translator training have existed long since the advent of MT some decades ago. GenAI has only caught the attention of researchers for around two years, so the number of studies is naturally still limited. While empirical studies on the integration of GenAI in translator training remain limited in number, existing research on MT has already shown the advantages and disadvantages of incorporating automatic translation in training (e.g., Doherty & Kenny, 2014; Zhang & Qian, 2023). Given that students are likely to independently explore and experiment with GenAI, just as they did with MT (Zhang, 2023), it is more beneficial to openly discuss these tools rather than prohibiting discussion and access in the translation classroom.

We thus argue that it would be better to understand how students have been using GenAI in translation tasks and provide tailored and essential guidance for them to leverage these tools. The first step in providing such tailored instructions is understanding students' usage of GenAI tools.

3 Prompt engineering

While empirical studies on students' interaction with GenAI are scarce, prompt engineering has emerged as a specialised technique applied across other fields, such as computational linguistics, healthcare and education (Mabrito, 2024; Patil et al., 2024; Reddy et al., 2024). This technique involves designing, refining, and implementing prompts (i.e., human input instructions) to optimise the output of GenAI to generate more accurate and contextually appropriate responses (Knoth et al., 2024; Ratnayake & Wang, 2024). Prompt engineering frameworks have gradually emerged

to guide practice. For example, the PERFECT Framework focuses on key elements, including prioritising Precision to reduce ambiguity, Engagement to make prompts relevant, Relevance to align with the task, Flexibility to allow varied responses, Efficiency to optimise resources, Clarity for understanding, and iterative Testing to refine prompts (Ratnayake & Wang, 2024). However, such studies rarely focus on translation-specific challenges.

Recently, the knowledge of prompt engineering has been transferred to and explored in the translation field by comparing translation quality: Studies that compare zero-shot and few-shot strategies (i.e., providing GenAI with no examples or a small number of examples to guide its response) have primarily focused on sentence-level translation and often overlooked the context (e.g., Hendy et al., 2023; Vilar et al., 2023); The level of input text have been considered in studies showing that full-document input yields better translation quality than sentence-by-sentence or multi-sentence block input (e.g., Wang et al., 2023), but the prompting strategies examined do not apply to real-world translation practice that considers functionalist principles, such as target audience and translation purpose (Vermeer & Chesterman, 2021). To our knowledge, only one study has provided a human-like prompting framework for translation, which includes four key components: M for Maps (keywords and terms), A for Audience (tone and style), P for Purpose (goal and context), and S for Style (maintaining consistency and cultural adaptation (He et al., 2024). However, this framework does not provide clear definitions of these translation terms, and it appears to be derived from experiential insights rather than from translation practice or established theoretical frameworks. As such, its potential applicability to professional translation contexts calls for further exploration and validation in authentic translation settings.

Against this backdrop, there is a need for a more systematic framework that is grounded in real-world translation practice and supported by empirical data, whether for guiding Human-GenAI translation practice or students in using GenAI in an informed manner. This study, therefore, aims to understand translation students' usage of GenAI tools by analysing their associated prompts.

We intend to address the following research questions (RQs):

- RQ1: What translation tasks are outsourced to GenAI tools by translation students?
- RQ2: What are the language features of the prompts used by translation students?
- RQ3: What are translation students' prompt engineering strategies?

To answer these RQs, we recruited 15 postgraduate students and collected their dialogues with GenAI tools over eight months for thematic and discourse analysis. The potential significance of this research lies in two key areas. Firstly, the findings of this research are expected to provide empirical evidence regarding how translation students interact with GenAI, particularly how they formulate and use prompts. Secondly, from a practical perspective, the findings could inform the development of effective pedagogical approaches for integrating GenAI into translator education.

4 Methodology

4.1 Data collection

After obtaining ethical approval from our institutions (Approval-No. 45644), we sent out a call to postgraduate students enrolled in a translation program jointly established by an Australian university and a Chinese university. Potential participants voluntarily contacted the research team to register their interest, and we asked them several follow-up questions to verify their eligibility. Eligible participants of the current project are students enrolled in translation programmes who have constantly experimented with GenAI tools to assist with their translation tasks, including real-life translation tasks and course assignments. Prior to this study, participants had neither received formal training in translation technology nor been permitted to use GenAI in their coursework. The prompts were created during the course as part of their regular learning activities, without participants being aware that these would later be collected for research purposes. Data collection began only after the coursework had concluded. Once the participants' eligibility was confirmed, they were given detailed instructions on exporting their dialogues created during translation tasks directly from the GenAI platforms and saving the dialogues in Word format. Participants were instructed to anonymise the files

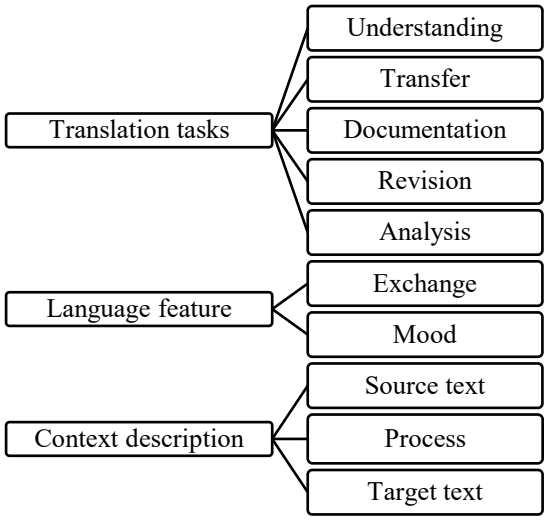


Figure 1: Coding typologies

by naming them with their assigned participant codes before uploading them to a shared Google Drive folder.

Fifteen participants were recruited, and most submitted eight documents spanning the eight months of the two terms of the 2024 academic year. The total number of prompts collected was 983 (excluding those unrelated to translation tasks) in 119 documents.

All the documents were imported into NVivo for further analysis. To improve the validity and reliability of the thematic and discourse analysis, the research team conducted the coding processes twice in December 2024 and January 2025. The results were compared to identify discrepancies, which were discussed among the research team members to reach a final decision.

4.2 Analytical framework

The data were analysed from three aspects: the use of GenAI in different translation tasks, the language features of the prompts used by students, and the description of the context provided by students. As displayed in Figure 1, The coding typologies were determined by observing our data and referencing relevant studies.

Regarding the use of GenAI in different translation tasks, we employed Mossop's (2000, p. 40) framework of three translation phases: pre-drafting, drafting (sentence-by-sentence drafting) and post-drafting. Five tasks were performed in these three phases, as follows: (1) Interpret the source text; (2) Compose the translation; (3) Conduct the research needed for Tasks 1 and 2; (4) Check the draft translation for errors and correct if necessary; (5) Decide the implications of the commission: how do the intended users and uses of the finished products affect Tasks 1 to 4?

For easier and clearer coding, the five tasks were indicated as understanding, transfer, documentation, revision and analysis.

Our discourse analysis of prompts drew upon the framework of dialogue analysis within Systemic Functional Linguistics (Halliday & Matthiessen, 2013). Considering that translation students interacted with GenAI tools following a dialogic structure (Batubara et al., 2024), this analytical approach focuses on the functional roles that language plays in communication and allows for a deep dive into the intentions behind exchanges. More specifically, we analysed prompts as individual utterances within the context of a dialogue framework, examining exchange patterns, interaction style, and utterance mood (i.e., linguistic features that reveal the speaker's attitude toward the action or state described in the sentence). The prompts were categorised according to three main types of mood: declarative (statement), interrogative (question), and imperative (command). Within each mood type, we further differentiated language functions based on the syntactic structure and word choice that manifest the interlocutor's different intentions in communication.

Regarding context descriptions in prompts, we employed a hybrid approach (Fereday & Muir-Cochrane, 2006), starting with open coding to capture any emerging themes in the prompts. Later, during the categorisation phase, we observed that some of the codes were closely related to existing translation frameworks. For example, codes relevant to textual functions of the Source text (ST) and Target text (TT) were interpreted within Snell-Hornby's integrated approach, which defines the domain (e.g., medical and legal text), genre (e.g., annual report and contract), audience (general or domain experts such as medical specialists) and other factors related to the communicative function of the text (Nord, 2018); Codes relevant to expected translation quality were referred to NAATI's models for assessing translation quality that involves transfer competency and language competency: Transfer Competency focuses on meaning transfer and adherence to textual norms, while language competency assesses the use of grammar, syntax, and idiomatic expressions to

ensure the translation is both accurate and appropriate for the target audience (NAATI, 2024). The coding system, therefore, integrated both existing theoretical frameworks and new insights derived from our data.

5 Results and discussion

5.1 GenAI in the translation processes

In examining how the participants utilised GenAI tools to assist with translation, the interaction evidently occurs in all five translation tasks across the translation process. Among these tasks, the *transfer* (59.86%) and *revision* (30.05%) tasks appeared to involve the most frequent and intensive use of GenAI. When transferring the ST into the TT, the participants often relied on GenAI to produce an initial translation draft for the entire text or some particularly challenging paragraphs. Some representative examples of prompts are listed as follows:

(1) Please translate the following text into English that aligns with natural English expressions. (P01)

(2) I have a document; could you please translate it? Keep the translation concise and elegant, with a literary style. (P06)

(3) Please help me creatively translate the following passage. (P06)

Revision also occurred mainly at the text or paragraph level and less frequently at the sentence or phrase level. The revision aimed to identify and correct translation errors by comparing the ST and TT, address awkwardness and ambiguity, and correct grammatical and syntactic errors in the TT. Several typical prompts were identified, as follows:

(4) Please polish and improve the translation so that it meets the requirements of English writing. (P01)

(5) Point out the errors of this translation. (P02)

(6) Can you rewrite one more time the translation. No need to make a lot of changes. Only need to correct the translations of some terms, grammar mistakes, and non-fluent sentences. Also make the translation more formal. (P06)

Though GenAI was less tasked with translating or revising a single sentence or phrase, the interaction in these cases tends to be more dynamic, often involving multiple dialogue exchanges. The participants frequently adjusted their prompts to ensure the output aligned with their desired style or quality. In contrast, a simple back-and-forth interaction was involved when translating or revising an entire text or paragraph, with one single prompt followed by GenAI's response. A representative example is presented below:

(7) Prompt 1: [An English sentence]. How should this sentence be translated in medical translation?

Prompt 2: How can the translation read more professionally?

Prompt 3: [part of GenAI's translation]. How can you say this differently?

Prompt 4: What if the translation has to sound more professional?

Prompt 5: It is still not fluent.

Prompt 6: Can you change the word order? (P06)

In example (7), the participant had one sentence translated by GenAI and was unsatisfied with the output because of the style. The participant then requested that the translation be revised to sound more professional. The participant also asked GenAI to provide a different version to choose from.

Another interesting observation is the preference for re-translation over revision. When the generated translation did not meet the expectations of the participants, a request to re-translate rather than revise the generated output was given with an updated prompt.

(8) Prompt 1: Please translate the following introduction of a medical company into English.

Prompt 2: [A paragraph from the ST]. Translate this paragraph again using four-character structures.

Prompt 3: [Two subtitles from ST]. Translate these two subtitles more elegantly.

Prompt 4: [Company brand name]. How can this brand name be translated into Chinese? (P02)

In example (8), the participant asked GenAI to translate an introduction to a medical company. The follow-up prompts all focused on re-translating some parts of the ST with updated instructions.

The application of GenAI is less significant in *terms of understanding (2.75%), documentation (5.5%),* and *analysis* (1.84%) tasks. The participants often employ GenAI to facilitate their *understanding* of the ST by asking it to provide a summary of the ST or to analyse the structure of some difficult sentences.

(9) Please read the readings and grasp some core ideas. (P01)

(10) Please analyse the sentence structure of the following sentence. (P10)

Regarding *documentation*, the participants prompted GenAI to explain domain-specific terms, proper names, or background information.

(11) What is the difference between [Term A] and [Term B]? (P12)

(12) Please help me compare and analyse the brand tones of [Brand A] and [Brand B] and present the comparison in a table format. (P05)

Concerning analysis, the participants required GenAI to help determine translation strategies.

(13) Please help me conduct a pre-translation analysis of this text.

The participants' interaction with GenAI in various translation tasks, on the one hand, highlights the multifaceted role of GenAI in translation workflows. As translators' roles may increasingly involve collaboration with GenAI tools, it is worth exploring the critical and creative application of GenAI throughout the entire translation process. More attention could be given to the tasks of *understanding*, *documentation* and *analysis*. On the other hand, such interaction with GenAI demonstrates that even without proper training, the participants have been experimenting with it and exploring its usage independently.

Several significant and interesting issues were revealed in our data. First, students' frequent application of GenAI in *transfer* and *revision* tasks shows its potential to accelerate translation processes by providing references. However, what matters is how students make use of the generated output, which requires further exploration. Second, fewer prompts directed toward the *understanding* and *documentation* tasks, in our opinion, may indicate students' reduced effort to double-check the generated translations, which means students' (potentially blind) trust of and (over-)reliance on GenAI. Third, as these participants have heavily engaged with GenAI, ethical issues should be discussed in the classroom, including intellectual property, transparency, and accountability.

5.2 Discourse features of the prompts

Our discourse analysis identifies structures and communicative functions of prompts to deepen our understanding of how translation students construct prompts through different language uses.

At the conversation level, the participants' prompts exhibit varying levels of interactivity when engaging with GenAI. 175 out of 356 conversations (49.16%) were limited to a single round, where the student commanded GenAI to translate a text, and GenAI's translated text marked the end of the exchange. In contrast, around half of the conversations between the participants and GenAI involved multiple rounds of exchanges with a continuous flow of information, responses, and feedback. In these multi-round exchanges, some prompts were context-dependent, lacking complete syntactic structure but were understandable within the given context (e.g. 'make it [the text] more logical' with the text provided in the previous prompt). Table 1 displays single-round and multi-round conversations between the participants and GenAI tools.

Single-round conversation	**Multi-round conversation**
<Beginning of conversation> Prompt: Translate into English. [The ST] GenAI output: [The TT] <End of conversation> (P04)	<Beginning of conversation> Prompt 1: I have a document; could you please translate it? Keep the translation concise and elegant, with a literary style. [The ST] GenAI output 1: [The TT] Prompt 2: How can [one phrase of the ST] be translated in a more literary way? GenAI output 2: [Suggest a different translation] Prompt 3: How to translate [a brand name] in a more appropriate way? GenAI output 3: [Analyse the brand name and point out the factors to consider when translating it] … <End of conversation> (P06)

Table 1: Examples of Single-Round and Multi-Round Conversations with GenAI Tools

Interestingly, increased interactivity was observed when AI was used to assist in examination tasks that contribute to final grades, while single-round conversations were primarily seen in weekly exercises. It remains inconclusive whether this difference is related to students' motivation; further observation of student-AI collaborative output or interviews with students will be needed to draw a definitive conclusion.

In addition to interactivity, we also identified informality of conversational language in the

Mood and function	Example
Imperative mood	
Command	*Refine* the above text, making the language more elegant, but avoid being overly verbose.
Request	*Please* translate the following text into English, following English expression conventions.
Assume	*Imagine* you are a medical translator who is translating the following text into English to make it fit for the needs of foreign patients and their families.
Suggest	*Consider* dividing this paragraph into four sections based on its logical structure to enhance readability.
Interrogative mood	
Confirm	*Do* these paragraphs have any linguistic mistakes or logic mistakes needed to fix*?* *Is there* any grammatical issue with this topic*?*
Request	*Can you help* me to translate?
Inquire	*How to* translate "population risk" into Chinese? What is IPG?
Critique	now, assume you are a native english speaker who has little idea about tibet and ways to travel to tibet, *are you interested to travel to tibet by railway after seeing the direct translation?*
Decide	*Does* 'limited access' mean they have difficulty obtaining it, *or* that the help they receive is limited*?*
Declarative mood	
Describe	It *is* a brochure and 13 20 50 is a telephone number.
Commissive	I *will* give you a picture for reference.
Evaluate	Some of your expressions are *hard* to understand for Chinese.
Explain	It needs to be simple and plain, *because* patients are busy with their own stuff. They need to catch the main information quickly.
Permit	You *can* add images to make it more like a brochure to attract people to the screening.
Permit	You *can* add images to make it more like a brochure to attract people to the screening.

Table 2: Moods and functions of prompts

prompts created by the participants: First, participants sometimes combined English and Chinese as the input language, for instance, 'justify 修改的部分，最好能够附上参考的 parallel texts (Justify the modifications made, better to include parallel texts as references) (P08)'. This reflects the phenomenon of bilinguals mixing languages in everyday communication (Ritchie & Bhatia, 2012). Second, the prompts contained typographical errors (e.g., 'into Chines') and grammatical mistakes (e.g., 'make some specific example about the translation'). Furthermore, colloquial expressions were present, such as '文邹邹' (wén zōu zōu), a misspelling of '文绉绉' (wén zhōu zhōu) that describes a style of speech or writing that is overly formal and pretentious (P05).

Following the analysis of the overall conversation structure and style, we further examined the prompts as individual utterances created by the participants, as shown in Table 2.

The conversational analysis of prompts revealed that the imperative mood was the most prevalent, particularly through its command function, which was used to instruct GenAI to perform translation tasks. This mood also encompassed requests, assumptions, and suggestions, characterised by action-oriented language that omits the subject and focuses on prompting specific actions. In addition, the study found that students also employed the interrogative mood when interacting with GenAI to seek clarification, validation, or new information. Such utterances typically featured question words or auxiliary verbs, reflecting the participants' need to engage with ChatGPT for further elaboration or problem-solving. The declarative mood was also used to convey information, express evaluations, explain reasoning, or grant permission. It was characterised by complete statements that provided factual, evaluative, or explanatory content, supporting the clear communication of ideas.

Unlike previous studies that focused on the content of prompts (e.g., He et al., 2024; Ratnayake & Wang, 2024), this study contributes by identifying and categorising the discursive features of prompts in terms of mood and communicative functions. This approach provides insights into the interactional patterns of translation students as both initiators and drivers of dialogue with GenAI tools. The findings also have potential implications for future training of GenAI models with analysing AI-generated products, as the categorisation of discursive features can inform the development of

Codes	Example
Author	The author is *a professor at an American university and a prominent left-wing feminist.* (P08)
Domain	Now translate a *medical* paper into Chinese. (P01)
Genre	Please help me translate the following material. It is *the annual report* of an agricultural development company. (P01)
Source	Below are the lyrics sung by a monk in *an English fictional novel*. Translate the lyrics into Chinese: (P04)
Theme	Please help me translate the following excerpt. It is about *the background information of the 'Belt and Road Initiative'.* (P05)
Text function	Translate the ST into Chinese…note that it's *a promotional material.* (P11)
Contextual background	The background information of this passage is: *In recent years, the growing wealth gap and political polarisation in the United States have led to increasing domestic doubts about this argument.* (P12)
Surrounding text	The function of "facilitators" in the sentence: *We have also collaborated with* facilitators *to help farmers create a "family vision plan," which focuses on tackling gender inequality and improving young people's access to …* (P01)

Table 3: Examples of prompts about ST background information

systems capable of recognising and responding to different prompt moods and communicative functions.

5.3 Context in prompt engineering

The open-ended thematic analysis was conducted to identify the contextual components that the participants used to craft prompts. The results revealed that 40.39% of the prompts (397 out of 983) only presented the text for processing and indicated the action (e.g., 'to translate' or 'to proofread') without providing any contextual information. For example:

Code	Example
Role	Assume you are *a medical translator.* (P01)
Application of knowledge from translation studies	
Theories	I need more examples *from the Skopos Theory.* (P10)
Approach	Re-translate, what does this mean? You may use *free translation* if appropriate. (P01)
Strategy	[ST in Chinese] How to translate this sentence? I need you to *explain* 吃得饱 and 吃得好 to English native speaker. (P06)

Table 4: Examples of prompts related to the translation process

(14) Translate into Chinese: [A sentence of the ST] (P01)

On the other hand, the prompts incorporating contextual information are relevant to *background information* about the ST, *requirements for the translation process*, and *expectations for the TT*.

Background information about the ST included components such as the **author** who has created the ST, the **domain** that specifies the field in which the ST is situated (e.g., medical, legal, or business domains), the **genre** (the type or category of the text, which shapes its structure and style), the **source** from which the ST is extracted, and **theme** of the ST. It also covered the **textual function** of the ST, the **contextual information** that involves the circumstances or environment in which the text was created, and the **surrounding text** located immediately before and after the ST. Representative examples are provided in Table 3.

The analysis also reveals themes that are relevant to *the translation process* (see examples in Table 4).

One key theme was the **role** assigned to GenAI tools, where prompts instructed the tools to adopt specific professional perspectives (e.g., assuming the role as a medical translator). We also observed that some of the prompts applied knowledge from **translation studies**, including theories, approaches, and strategies.

In addition, participants provided examples in their prompts to guide GenAI's responses, such as providing translated text that can be used in the generated output and specifying writing styles for GenAI to reference.

Code	Example
Domain of the target text	Please use *legal* language. (P09)
Genre of the target text	How to express this in *an academic paper*.
Audiences who intend to read the translated text	Need to be presented to *Chinese medical researchers*. (P01)
Text function	You have been asked to translate the following *for marketing the product …* (P01)
Format	
Syntactic structure	Turn the above content into *a dialogue format* for communication with the translation company. (P06)
Length	Shorten the answer, *no more than 250 words.* (P15)
Expected quality standards	
Accuracy	Please help me translate the following sentences into English, with a focus on *fidelity and accuracy.* (P05)
Application of textual norms and conventions	
Writing style of the target text	*Use more common language* to explain some professional terms. (P12)
Use of **terminology**	The passage serves as a parallel text, based on this, plz polish your answer, especially the *terms*, make sure your translation is accurate. (P11)
Language quality of the translated text	
Idiomatic expressions	Please translate the following into English, adhering to *English expression conventions.* (P01)
Grammar	The ST consists mostly of subjectless sentences. Please ensure to *add subjects* in the translation. (P12)
Coherence and **cohesion**	Polish the paragraph, make it more *cohesive and coherent* and appealing. (P11)

Table 5: Examples of prompts related to expectations on the TT

(15) *…9.shall timely report the relevant information to 10. the public security department* Replace with these terms, and generate another translation version. (P07)

(16) Translate the following English text according to the style of the given Chinese translation. *[Source text in English][Example in Chinese]* (P08)

Regarding *the output generated by GenAI,* the participants mentioned information relevant to their expectations on the TT in their prompts. Examples are presented in Table 5.

Our analysis reveals that participants consciously included information in their prompts about the expected textual features, functions, formatting and quality standards (including accuracy of meaning transfer, the appropriate application of textual norms, and overall language quality). However, their descriptions of translation quality often relied on abstract words that are not clearly defined and may be interpreted differently by different people, for example, 'translate it more elegantly' (P02), 'more attractive' (P04), and 'more idiomatic' (P15). Using such words may introduce ambiguity that results in non-expected responses from GenAI.

In summary, the results indicate that a considerable number of prompts lacked specific task descriptions. This may potentially limit GenAI's ability to generate accurate translations, as previous studies have highlighted the inclusion of contextual components in prompts as an effective strategy for improving AI-generated results (e.g., Park & Choo, 2024; Ratnayake & Wang, 2024). Approximately half of the participants consciously included descriptions of the translation process and quality expectations in their prompts. While these prompt strategies were often unsystematic and characterised by ambiguous or abstract descriptions, they nonetheless demonstrated the incorporation of translation-specific knowledge. The prompts show discipline-driven deviations from general prompt engineering strategies (e.g., He et al., 2024; Hendy et al., 2023) that echo approaches from descriptive translation studies (e.g., Nord, 2018). As GenAI development increasingly shifts toward task-specific solutions (Yehia, 2024), these findings not only help identify students' intuitive prompting behaviours and gaps before training, informing translation-specific GenAI instruction, but also offer insights for future research on refining GenAI functionalities to better support translation practice.

6 Concluding remarks

To answer the research questions posed in the current study, we collected and analysed student participants' prompts to explore their interaction with GenAI in translation tasks. Our findings revealed that the student participants interacted

with GenAI across various tasks, especially transfer and revision, in the translation process, even without proper training. In terms of discourse features, it is common that student participants' interactions with AI ended after a single round, lacking necessary iterative feedback and refinement, with prompts reflecting an informal, spoken language style. Analysis of sentence structures and word choices further revealed the student participants' diverse prompting strategies, shaped by their language use. Regarding the content of the prompts, the findings indicate a lack of awareness in incorporating contextual cues, which may limit the effectiveness of GenAI in generating appropriate translations. It was evident that the student participants applied translation theories to their prompts, demonstrating an understanding of translation concepts and quality criteria; however, their use of vague, abstract terms may introduce ambiguity, leading to less accurate AI outputs. Overall, these interactions provide valuable insights into how GenAI can be integrated to improve educational interventions and professional practice. Our findings can serve as references for designing specialised prompt engineering training for translation students, practitioners' professional development, and future studies analysing the products of student–GenAI interactions. Our findings also suggest that these future translators increasingly rely on human-AI collaboration, thus posing new challenges for educators to urgently review translation education and adapt to this rapidly evolving landscape.

References

Muhammad Hasyimsyah Batubara, Awal Kurnia Putra Nasution, and Fachrur Rizha. 2024. ChatGPT in communication: A systematic literature review. *Applied Computer Science*, 20(3):96–115.

Stephen Doherty. 2016. The impact of translation technologies on the process and product of translation. *International Journal of Communication*, 10:947–969.

Stephen Doherty and Dorothy Kenny. 2014. The design and evaluation of a Statistical Machine Translation syllabus for translation students. *Interpreter and Translator Trainer*, 8(2):295–315.

Stephen Doherty and Joss Moorkens. 2013. Investigating the experience of translation technology labs: pedagogical implications. *Journal of Specialised Translation*, 19:122–136.

Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, 5(1):80–92.

Linling Fu and Lei Liu. 2024. What are the differences? A comparative study of generative artificial intelligence translation and human translation of scientific texts. *Humanities and Social Sciences Communications*, 11(1):1–12.

Michael Alexander Kirkwood Halliday and Christian MIM Matthiessen. 2013. *Halliday's introduction to functional grammar*. Routledge.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How GPT Models at A arXiv:2302.09210 [cs].

Miguel A. Jiménez-Crespo. 2024. Transcreation in and the of AI: Focusing on " In Loukia Kostopoulou and Parthena Charalampidou, editors, *New Perspectives in Media Translation*, pages 309–320. Springer International Publishing, Cham.

David Katan. 2022. Tools for transforming translators into homo narrans or "what machines can't do." In *The Human Translator in the 2020s*, pages 74–90. Routledge.

Dorothy Kenny and Stephen Doherty. 2014. Statistical machine translation in the translation curriculum: overcoming obstacles and empowering translators. *Interpreter and Translator Trainer*, 8(2):276–294.

Nils Knoth, Antonia Tolzin, Andreas Janson, and Jan Marco Leimeister. 2024. AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 6:100225.

Tong King Lee. 2023. Artificial intelligence and posthumanist translation: ChatGPT versus the translator. *Applied Linguistics Review*, 15(6):2351–2372.

Fangyuan Li and Lu Tian. 2024. Translation practice and competence enhancement in the age of AI: Applying ChatGPT to translation education. In *Lecture Notes in Computer Science*, volume 14606 LNCS, pages 219–230. Springer.

Fengqi Li, Zhijian Cao, and Xinchen Li. 2023. College translation teaching in the era of artificial intelligence: Challenges and solutions. *Journal of*

Higher Education Theory and Practice, 23(19):39–49.

Marek Wojciech Łukasik. 2024. The future of the translation profession in the era of artificial intelligence: Survey results from Polish translators, translation trainers, and students of translation. *Lublin Studies in Modern Languages and Literature*, 48(3):25–39.

Mark Mabrito. 2024. Artificial intelligence in the classroom: Conversation design and prompt engineering for English majors. *International Journal of Technologies in Learning*, 31(2).

McKinsey & Company. 2023. The economic potential of generative AI: The next productivity frontier. Technical report. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#introduction. Accessed April 04, 2025.

Sahar Yousif Mohammed, Abed Shahooth Khalaf, Mohammed Aljanabi, and Maad M. Mijwil. 2024. Challenges and opportunities in translation studies: The evolving role of Generative AI in translation development. In Nadia Mansour and Lorenzo M. Bujosa Vadell, editors, *Sustainability and Financial Services in the Digital Age*, pages 107–117. Springer, Cham.

Mohammed Mohsen. 2024. Artificial intelligence in academic translation: A comparative study of large language models and Google Translate. *Psycholinguistics*, 35(2):134–156.

Brian Mossop. 2000. The workplace procedures of professional translators. In Andrew Chesterman, Natividad Gallardo San Salvador, and Yves Gambier, editors, *Translation in Context: Selected papers from the EST Congress, Granada 1998*, volume 39 of *Benjamins Translation Library*, pages 39–48. Benjamins.

NAATI. 2024. NAATI assessment rubrics. https://www.naati.com.au/news/naati-releases-refined-assessment-rubrics-on-1-april-2024/. Accessed April 04, 2025.

Christiane Nord. 2018. *Translating as a purposeful activity: Functionalist approaches explained*. Routledge, 2nd ed.

Jiyeon Park and Sam Choo. 2024. Generative AI prompt engineering for educators: Practical strategies. *Journal of Special Education Technology*:01626434241298954.

Rajvardhan Patil, Thomas F. Heston, and Vijay Bhuse. 2024. Prompt engineering in healthcare. *Electronics*, 13(15):2961.

Huaqing Wu, Lenny Yang, Arthur Wan, and Ming Qian. 2024. Augmented machine translation enabled by GPT4: Performance evaluation on human-machine teaming approaches. In *Proceedings of the First Workshop on NLP tools and Resources for Translation and Interpreting Applications.*

Himath Ratnayake and Can Wang. 2024. A prompting framework to enhance language model output. In Tongliang Liu, Geoff Webb, Lin Yue, and Dadong Wang, editors, *AI 2023: Advances in Artificial Intelligence*, volume 14472 of *Lecture Notes in Computer Science*, pages 66–81. Springer, Singapore.

C. Kishor Kumar Reddy, Pellate Anoushka, Akhil Draksharapu, and Srinath Doss. 2024. Beyond Text: Analyzing artificial intelligence models through prompt engineering. In Inam Ullah Khan, Hamed Taherdoost, Mitra Madanchian, Ouaissa, Salma EI Hajjami, and Hameedur Rahman, editors, *Future Tech Startups and Innovation in the Age of AI*, pages 120–156. CRC Press. publisher: CRC Press.

William C. Ritchie and Tej K. Bhatia. 2012. Social and psychological factors in language mixing. In Tej K. Bhatia and William C. Ritchie, editors, *The Handbook of Bilingualism and Multilingualism*, pages 375–390. Wiley, 1st ed.

Yousef Sahari, Abdu M. Talib Al-Kadi, and Jamal Kaid Mohammed Ali. 2023. A cross-sectional study of ChatGPT in Translation: Magnitude of use, attitudes, and uncertainties. *Journal of Psycholinguistics Research*, 52(6):2937–2954.

Kizito Tekwa. 2024. Artificial intelligence, corpora, and translation studies. In Defeng Li and John Corbett, editors, *The Routledge Handbook of Corpus Translation Studies*, pages 103–118. Routledge.

Hans J. Vermeer and Andrew Chesterman. 2021. Skopos and commission in translational action. In Lawrence Venuti, editor, *The translation studies reader*, pages 219–230. Routledge.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. arXiv:2211.09102 [cs].

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. arXiv:2304.02210 [cs].

Engy Yehia. 2024. Developments on Generative AI. In Purvi Pokhariyal, Archana Patel and Shubham Pandey, editors, *AI and emerging technologies: Automated decision-making, digital forensics, and ethical considerations*, pages 139–160. Routledge.

Jia Zhang. 2023. Exploring undergraduate translation students' perceptions towards machine translation:

A qualitative questionnaire survey. In Masaru Yamada and Felix do Carmo, editors, *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 1–10. Asia-Pacific Association for Machine Translation.

Jia Zhang. 2025. Too tricky for rookies? An enquiry into novice translation students' machine translation literacy. In Song Ge and Chen Xuemei, editors, *Multilingual Education Yearbook 2025 - Translation Practices as Agents of Transformation in Multilingual Settings*. Springer.

Jia Zhang and Hong Qian. 2023. The impact of machine translation on the translation quality of undergraduate translation students. In Masaru Yamada and Felix do Carmo, editors, *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 99–108. Asia-Pacific Association for Machine Translation.

Wenjuan Zhao, Siyu Huang, and Lizhen Yan. 2024. ChatGPT and the future of translators: Overview of the application of interactive AI in English translation teaching. In *2024 4th International Conference on Computer Communication and Artificial Intelligence (CCAI)*, pages 303–307, Xi'an, China. IEEE.

Can postgraduate translation students identify machine-generated text?

Michael Farrell
IULM University
Milan
Italy
michael.farrell@iulm.it

Abstract

Given the growing use of generative artificial intelligence as a tool for creating multilingual content and bypassing both machine and *traditional* translation methods, this study explores the ability of linguistically trained individuals to discern machine-generated output from human-written text (HT). After brief training sessions on the textual anomalies typically found in synthetic text (ST), twenty-three postgraduate translation students analysed excerpts of Italian prose and assigned likelihood scores to indicate whether they believed they were human-written or AI-generated (ChatGPT-4o). The results show that, on average, the students struggled to distinguish between HT and ST, with only two participants achieving notable accuracy. Closer analysis revealed that the students often identified the same textual anomalies in both HT and ST, although features such as low burstiness and self-contradiction were more frequently associated with ST. These findings suggest the need for improvements in the preparatory training. Moreover, the study raises questions about the necessity of editing synthetic text to make it sound more human-like and recommends further research to determine whether AI-generated text is already sufficiently *natural-sounding* not to require further refinement.

1 Introduction

Authors writing in a second language can today bypass the traditional process of writing in their native language and then having their work translated – either by a human translator or through machine translation – by engineering customized prompts for generative artificial intelligence (GenAI). These prompts, which may be written in the author's native language, the target language or a combination of both, include a precise description, outline or rough draft of the intended text. Consequently, there is no source language document in the traditional sense.

Content generated in this way may then be refined by a human synthetic-text editor tasked with enhancing its engagement and giving it a more human-like tone. This type of editing requires a skill set distinct from that used in post-editing, as the textual anomalies present in synthetic text (ST), – such as redundancy, blandness, verbosity, low burstiness and lack of complex analysis – differ from those typically seen in raw machine translation output (Dou et al. 2022; Farrell, 2025a).

These anomalies appear to be potentially language-independent. For example, redundancy – defined as the repetition of information without adding new meaning or value – can occur in texts written in any language.

The need for synthetic-text editing (STE) assumes that readers are indeed capable of distinguishing AI-generated output from human-written text (HT). Moreover, the ability to identify the textual anomalies characteristic of ST is essential for effective STE.

Clark et al. (2021) observed that untrained, non-expert evaluators are not well equipped to detect machine-generated English text, and even with training, their detection success rate improved only slightly, reaching about 55%. In their study, the evaluators were recruited through Amazon Mechanical Turk and were screened only by location/language (English) and their approval rating on the platform. They did not possess specialized knowledge, such as familiarity with Large Language Models (LLMs) or a background

Proceedings of Machine Translation Summit XX Volume 1, pages 432–441, June 23-27, 2025

in linguistics[1]. Conversely, Dou et al. found that English ST and HT could be distinguished after laypeople (also recruited through Amazon Mechanical Turk) annotated the texts using a framework called *Scarecrow*, which defines specific error types.

2 Aims

The principal objective of this experiment was:

- To evaluate whether postgraduate translation students can effectively identify Italian ST after brief training sessions.

There were also several secondary aims:

- To have the students identify examples of textual anomalies that can be used to enhance the training material, and to determine whether the same categories of ST anomalies found in English texts also occur in Italian.
- To refine the training instructions by identifying areas that require clarification or adjustments to reduce the occurrence of false positives.
- To shed light on the need for STE. If postgraduate translation students cannot reliably distinguish ST, it may already be sufficiently human-like without the need for further editing.
- To assess whether ChatGPT-4o can be guided through prompts to avoid the types of anomalies typically observed in ST.

3 Method

Twenty-three postgraduate translation students at the IULM University in Milan, Italy, attended two 30-minute lessons, held one week apart, introducing LLMs, generative artificial intelligence (GenAI) and some common anomalies reported in ST (Dou et al. 2022; Farrell, 2025a). During these lessons, a few examples of textual anomalies were provided, with the hope that the experiment itself would generate additional examples to improve future training materials.

The participants were then presented with 28 short excerpts (ranging from 268 to 467 words) drawn from seven Italian short stories, divided into four sets of seven excerpts each (A, B, C and D). They were informed that each set contained at least one HT and at least one ST excerpt. In reality, each set contained precisely one sequential excerpt of approximately equal length from Alberto Moravia's short story *L'incosciente* (The Reckless Man), from *Racconti romani* (Roman Tales, 1954), along with six sequential excerpts from unabridged short stories generated by ChatGPT-4o using prompts engineered as described below. The order of excerpts was randomized within each set.

The excerpt sets were assigned based on the students' seating arrangement in the lecture room. The student sitting in the first row on the right (from the lecturer's point of view) was assigned set A, the student to their right was assigned set B, the next student set C, and so on, cycling through the sets to ensure a roughly equal distribution. The students were instructed to move on to the next alphabetical set if they finished evaluating their initial set before the allotted time expired. The participants working on set D were instructed to proceed to set A. The experiment concluded once the researcher judged that every student had analysed at least one complete set.

The students were asked to assign a score from 0 to 10 to each text excerpt based on its likelihood of being machine-generated (0 = human-written; 10 = machine-generated; 5 = uncertain). Intermediate integer scores were allowed. They were also asked to identify and classify the types of anomalies or errors that influenced their assessments according to the categories illustrated during the training sessions. Due to time constraints, the participants were encouraged, but not required, to provide specific examples of the anomalies they identified.

To prevent the students from distinguishing the HT excerpts by finding them online, they were not allowed to consult the internet during the experiment. They were also not allowed to speak to other people, including fellow participants.

A few weeks later, a debriefing session was held, where the students were asked to provide feedback on the experiment and training through a preliminary questionnaire, a class discussion and a final questionnaire identical to the first, to determine whether the discussion had caused them to change their opinions.

[1] Unpublished clarification courtesy of Elizabeth Clark.

3.1 Prompt engineering

A prompt reverse-engineering approach, based on the Automatic Prompt Engineer technique (Zhou et al., 2022), was used because it effectively extracts the storyline from a story, allowing the AI-generated output to follow a similar narrative structure to the human-written one. The aim was to minimize the influence of subjective preferences regarding differing content or theme.

The initial prompt was generated by ChatGPT-4o itself by uploading Alberto Moravia's short story and entering the following instruction:

> *"If I had to write a prompt that would cause you to generate the Italian text in the attached file, what would it be? Keep in mind that it is 1808 words long, including the title."*

The first artificially generated story (ST1) was then generated by entering the prompt provided by ChatGPT-4o (Appendix A) into a new chat.

The second AI-generated story (ST2) was produced similarly but with modifications to the prompt to set the story in Rome and to name the young protagonists Emilio and Santina, as in Moravia's original. The following additional instruction was also appended to the new prompt:

> *"Machine generated text is often criticized for the excessive repetition of words or phrases; the repetition of information without adding new meaning or value; the absence of emotion, creativity and engagement; overly long, highly descriptive, fanciful sentences; uniform sentence structure and length; and lack of complex analysis. Make sure the generated text does not have any of these anomalies."*

For ST3, the prompt retained the same setting and character names but replaced the instruction to avoid textual anomalies with:

> *"Write the text in the style of the Italian author Alberto Moravia (1907–1990)."*

ST4's prompt shifted the setting to a neighbourhood on the outskirts of Naples and the protagonists were renamed Emilio Capuozzo (Mimì) and Santina Picariello (Tina). It also specified that the story should be written in the style of Italian author Elena Ferrante, whose Neapolitan Novels are set similarly.

ST5 was set in Asti, with protagonists Emilio and Santina, and was written in the style of Italian crime writer Giorgio Faletti, a native of Asti.

ST6 moved to Florence, again with Emilio and Santina as protagonists. The requested style was that of the Florentine journalist and author Oriana Fallaci. All six AI-generated stories (ST1–ST6) were produced on 31 August 2024.

In all cases, the prompts specified that the generated Italian short stories should be approximately 1800 words long. However, the AI-generated stories turned out to be shorter than Moravia's original (HT0). To ensure excerpts of comparable length, the last 271 words of HT0 were omitted. Each story was divided into four consecutive excerpts of approximately equal length, avoiding splits mid-paragraph, and one excerpt was placed into each of the four sets of seven (A, B, C and D).

Although ChatGPT-4o was asked to generate similar short stories to reduce the effect of subjective preferences for certain topics, stylistic variation was deliberately introduced by requesting different writing styles based on well-known Italian authors in order to avoid the AI-generated stories being identified due to their similarity.

Lastly, the ST stories and HT0 were analysed using Plagramme AI detector [2] to determine whether any objectively measurable differences existed between them.

4 Results

Twenty-three students took part in the experiment. Each one analysed an average of 7.74 text excerpts, with the number of assessments ranging from a minimum of 4 (by 1 participant) and 6 (by 2 participants) to a maximum of all 28 (by 1 participant). As shown in Table 1, on average, the students were unable to identify HT0 since they assigned it an overall mean score of 5.22 (indicating uncertainty). In fact, four of the six AI-generated stories were, on average, perceived as more *human-like* than HT0. None of the short stories were clearly identified as ST, with the highest overall mean score being 5.85 (still very close to uncertain).

However, two students (8.70% of the 23 participants), Student No. 8 and Student No. 20, showed a notable above-average ability to distinguish the HT from the ST excerpts.

Student No. 8 analysed a total of eight excerpts, consisting of all seven excerpts in set C and one excerpt from set A (HT0 A), even though she

[2] www.plagramme.com

Text	Length (words)	Mean Student score	AI Detector score
HT0 A	359	4.13	36%[a]
HT0 B	324	7.14	13%
HT0 C	386	5.43	16%
HT0 D	467	4.00	8%
Entire HT0	**1536**[b]	**5.22**[c]	**17%**
ST1 A	373	4.86	64%
ST1 B	420	6.33	100%
ST1 C	408	4.71	72%
ST1 D	389	4.33	89%
Entire ST1	**1590**	**5.04**	**86%**
ST2 A	333	1.86	97%
ST2 B	392	2.17	81%
ST2 C	338	4.86	94%
ST2 D	268	3.00	100%
Entire ST2	**1331**	**3.00**	**83%**
ST3 A	375	1.67	94%
ST3 B	398	3.67	86%
ST3 C	399	3.43	87%
ST3 D	376	5.67	89%
Entire ST3	**1548**	**3.60**	**85%**
ST4 A	373	6.43	65%
ST4 B	334	7.60	96%
ST4 C	367	4.57	94%
ST4 D	374	2.60	87%
Entire ST4	**1448**	**5.33**	**83%**
ST5 A	420	2.00	94%
ST5 B	379	4.00	93%
ST5 C	417	4.14	86%
ST5 D	410	3.40	82%
Entire ST5	**1626**	**3.36**	**94%**
ST6 A	306	7.29	84%
ST6 B	341	6.60	99%
ST6 C	339	5.29	89%
ST6 D	331	4.43	93%
Entire ST6	**1317**	**5.85**	**73%**

Table 1: Average scores assigned to the text excerpts by the students.

a) This excerpt includes a paragraph that received an anomalous score of 99% according to Plagramme AI detector.

b) To keep the excerpts to approximately the same length, the last 271 words were not used.

c) The overall mean in each case does not equal the mean of the partial means because the number of students evaluating each excerpt varies.

should have moved on to set D. She accurately assigned a score of 0 to both excerpts written by Alberto Moravia. For the remaining six ST excerpts, she gave scores ranging from 5 (uncertain) to 10 (definitely machine-generated).

Student No. 20, on the other hand, analysed the seven excerpts in set D. The text to which she gave the lowest score (3) – signifying it appeared the most human – was the only HT excerpt she assessed. She assigned relatively high scores, ranging from 7 to 9 (indicating somewhere between probably and almost definitely AI-generated) to the six ST excerpts.

If we exclude the excerpt from set A that Student No. 8 analysed, the two students become directly comparable, since they both evaluated a complete set of seven excerpts, each containing one HT.

Now, let's calculate the probability that these two students correctly identified the HT excerpt purely by chance. The participants were told that at least one excerpt in each set was human-written and at least one was AI-generated. Based on this, there are six possible scenarios per set, ranging from "only one excerpt is HT" to "six of the seven excerpts are HT". Hence, the probability of a student guessing that only one of the seven excerpts is written by a human is 1/6.

Assuming they correctly guess that there is only one HT excerpt, the probability of guessing which one it is without looking at them is 1/7, as each set contains seven excerpts. Since these two guesses are independent, the combined probability of making both guesses correctly is 1/6 * 1/7 = 1/42, or approximately 2.38%.

However, 23 students took part in the experiment, and two of them identified the HT excerpt. The probability that at least two participants out of 23 guess correctly without analysing the excerpts can be calculated using the binomial probability formula:

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Where n represents the total number of participants (n=23), k is the number of successful students (k=2), and p is the probability that an individual participant guesses correctly (p=1/42), the probability that at least two students out of 23

Anomaly	Text						
	H0	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
Excessive repetition of words or phrases	**13**	10	4	8	4	5	11
Redundancy	**9**	8	3	7	6	5	8
Non-existent words	1	0	0	0	0	**1**	0
Blandness	**10**	7	6	3	7	3	6
Verbosity	**11**	2	3	1	3	3	2
Low burstiness	4	9	5	9	**9**	8	8
Lack of complex analysis	7	**10**	7	5	6	3	5
Grammar and spelling mistakes	**11**	6	7	2	6	4	9
Hallucination	**4**	3	0	1	3	0	2
Self-contradiction	2	1	1	1	**3**	2	2
Unnecessary technical jargon	**4**	0	0	0	0	2	0
*Total replies from the students**	*27*	*26*	*25*	*25*	*24*	*25*	*26*

Table 2: Textual anomalies detected by the students by text. The highest scores are highlighted in bold red. *This number exceeds the total number of participants because some students analysed more than one excerpt from the same short story.

succeed purely by chance is approximately 10.32%[3].

This relatively low probability strongly suggests that the two students in question used analytical skills, rather than random guessing, to distinguish the AI-generated excerpts from the human-written ones during the experiment.

4.1 Textual anomalies detected

Table 2 clearly shows that the participants found most of the textual anomalies they were asked to detect in both the HT and ST excerpts. In fact, the human-written story was perceived as the most artificial text in 7 out of the 11 categories.

Despite this, the results support the assumption that the same ST anomaly categories observed in English ST also occur in Italian ST, with the possible exception of non-existent words, which were absent (see Section 6.2.1), and the notable exception of grammar and spelling mistakes. While such mistakes are relatively rare in artificially generated English texts (Dou et al., 2022; Gillham, 2024), they were found to be common in the Italian ST excerpts (see Section 6.2.3).

Owing to the time constraints mentioned earlier, not all the students provided specific examples of the anomalies they reported: 19 gave examples of grammar and spelling mistakes, 16 of excessive repetition of words or phrases, 11 of redundancy, 9 of low burstiness, 7 of verbosity, 7 of hallucination, 6 of self-contradiction, 6 of unnecessary technical jargon, 3 of lack of complex analysis, 2 of non-existent words (both spurious) and 2 of blandness.

The examples of burstiness and self-contradiction may be used in the future to enhance the training material (see Section 7).

4.2 Debriefing

Only eight students attended the debriefing session held a few weeks after the experiment. A ninth student joined later, but her replies were not analysed because she had not completed the initial questionnaire.

During the session, the students were shown the overall mean scores in Table 1 and asked to complete a closed-answer questionnaire on why so many of them had failed to distinguish between the ST and HT excerpts. This was followed by an open discussion covering the questionnaire topics, the experiment itself, the preparatory training and general observations about ST.

After the discussion, the participants were asked to complete the same questionnaire again, with exactly the same questions, to determine whether the classroom discussion had altered their opinions.

Half of the participants, including the only successful student present, stated that the textual anomalies they were asked to look out for could

[3] Using the Statology binomial distribution calculator: www.statology.org/binomial-distribution-calculator

also be found in HT0. Despite this, at the beginning of the session, three-quarters of the participants, including the successful student, disagreed with the hypothesis that searching for textual anomalies is an ineffective method for identifying ST. However, following the discussion, this proportion dropped to just over one-third (37.5%), although the successful student maintained her original position.

None of the students found the text excerpts too long, and the majority after discussion (62.5%) did not feel they needed to be longer.

Possibly due to a growing sense of disappointment, the percentage of the students who disagreed with the hypothesis that humans cannot distinguish between ST and HT, regardless of the training received, dropped from 75% at the beginning of the classroom discussion to 37.5% by the end. However, no one explicitly agreed with the proposition.

Following the classroom discussion, half of the students deemed the training insufficient and expressed the need for more practice.

The results for the most significant questions, both before and after the discussion, are shown in Appendix B (Table 3 to Table 8). The replies of the only successful student present, Student No. 20, are highlighted in bold red.

5 Limitations

Since this experiment was conducted as part of a postgraduate degree course with set number of hours, it was necessarily limited to a small selection of texts of a similar kind in a single language. The time available for preparatory training in GenAI detection was also limited. Moreover, the size of the class restricted the number of participants. As a result, the findings and conclusions of this study may not be broadly generalizable. However, the practical, hands-on learning experience and potential contribution to course development outweigh these limitations.

6 Discussion

6.1 Postgraduate translation students as evaluators

Judging from the results shown in Table 1 and Table 2, the answer to the question posed in the title of this paper (*Can postgraduate translation students identify machine-generated text?*) appears, at first glance, to be a resounding no.

This result is all the more disappointing considering that postgraduate translation students possess a background in linguistics, which might make them more qualified than the evaluators used in the two studies mentioned in the introduction (Clark et al., 2021; Dou et al., 2022).

In contrast, Plagramme AI detector showed little doubt in its assessments, assigning the ST stories probabilities of being AI-generated of between 73% and 94%, while attributing only a 17% likelihood to Alberto Moravia's work. Moreover, there was no alignment between the AI detector's scores and the mean scores given by the students.

However, closer analysis of individual participant data, as noted in Section 4, reveals that two out of the 23 students involved in the experiment are very probably able to distinguish ST from HT, at least as regards the specific texts analysed in this study.

6.2 Textual anomalies

As mentioned in Section 4.1, the students identified most of the textual anomalies they were asked to look out for in both the HT and ST excerpts. This finding was further confirmed during the debriefing session, as mentioned in Section 4.2. The following subsections provide a more detailed discussion of the results regarding specific anomalies.

6.2.1 Non-existent words

The student who reported non-existent words in the ST excerpts clarified in a note that she was not actually identifying non-existent words but rather pointing out the unusual use of certain terms. Similarly, the student who flagged a non-existent word in HT0 explained that she was referring to the French word *parabrise*, which – though uncommon – is occasionally used in Italian prose[4]. Neither of these cases involves truly non-existent words, like the term *grasitating* reported in an earlier experiment by Farrell (2025a).

It should be noted in fairness that the participants were not allowed to use a browser to ensure they could not identify which story was human-written by finding parts of it online. Consequently, they were unable to verify the existence of any unusual terms they encountered.

[4] www.treccani.it/enciclopedia/ricerca/parabrise/?search=parabrise

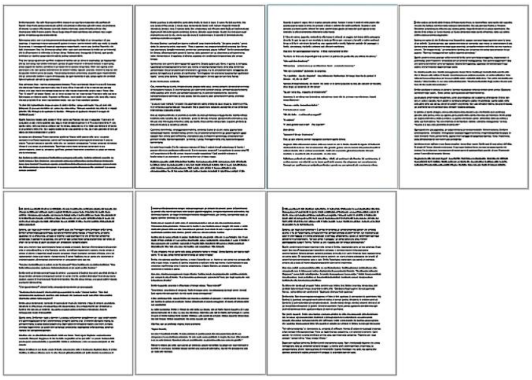


Figure 1: Thumbnails of the first page of each story in print layout view.

In any case, non-existent words could theoretically occur in HT as a result of typos, potentially leading to false positives (see Section 6.2.3).

6.2.2 Low burstiness

Burstiness measures variation in writing patterns, including sentence structure and length. Unlike machines, humans tend to exhibit high burstiness by naturally varying their writing to prevent repetition, such as by avoiding multiple sentences that start in the same way. Table 2 shows that most of the students who reported low burstiness correctly associated it with the ST excerpts. When Student No. 8 was asked how she had been so successful, she explained that, in her opinion, the key to identifying ST was noting the use of simple, very short sentences, adding that this brevity was clearly not intended for stylistic purposes.

These findings suggest that low burstiness was the most effective indicator in this experiment among the anomaly categories analysed. Notably, burstiness is also one of the parameters measured by AI detectors, such as GPTZero (Chaka, 2023).

Interestingly, it appears that the burstiness of the stories used in this specific experiment can be roughly estimated simply by examining their print layout, provided there is enough text. To test this idea, a small additional experiment was conducted with six randomly chosen undergraduate translation students from the same university. They were shown illegible thumbnails of the first page of the seven short stories used in the postgraduate experiment, presented in random order, and asked whether any of them stood out in terms of layout. All six students unequivocally indicated HT0 (the third from the left in the top row of Figure 1). It probably stands out due to its greater use of dialogue, which is also found to a lesser extent in the six ST stories. It would be useful to investigate whether this quick, simple detection method can be generalized to other texts, authors, genres and languages.

6.2.3 Grammar and spelling mistakes

Grammar and spelling mistakes are known to be relatively rare in artificially generated English texts (see Section 4.1). However, they are more common in AI-generated Italian texts. In this experiment, the students identified a few examples, including:

1. *Nei giorni seguenti, Emilio evitò Santina, temendo che lei potesse capire cosa* ***stava*** *succedendo.*[5]
   Correct Italian grammar requires the use of the subjunctive tense "…**stesse** succedendo".
2. *Santina lo fissò, sorpresa, ma Emilio non cercò* ***il suo*** *approvazione.*[6]
   The article and possessive adjective should agree with the noun "…**la sua** approvazione".

The students also identified grammar issues in HT0. However, it is likely that Moravia intentionally used unconventional grammar, such as *"per me, io ci sto"*.[7] as a stylistic device to reflect the social and cultural backgrounds of the characters in his stories, thereby adding authenticity. Indeed, he wrote the story used in this experiment *The Reckless Man* in the first person, imagining himself as a young working-class boy in post-war Rome. Moreover, typos are not uncommon in printed texts, meaning that an error like Example 2 above could also theoretically appear in HT.

Given these factors, grammatical accuracy and spelling seem to be highly unreliable parameters for distinguishing between Italian ST and HT.

6.2.4 Hallucination and self-contradiction

All but one of the instances of hallucination reported in HT0 were, in reality, unusual or antiquated turns of phrase (for instance, *non posi*

[5] Over the next few days, Emilio avoided Santina, fearing that she might realize what was going on.
[6] Santina stared at him in surprise, but Emilio didn't seek her approval.
[7] Count me in.

tempo in mezzo[8]). If the students had been allowed to consult the internet, they would probably have discovered that these expressions exist and might not have flagged them as hallucinations. The remaining example was *custode del passaggio a livello*. While referring to level-crossing guards may seem *hallucinatory* today, they did exist in Italy at the time when Moravia's story was set.

All the cases of hallucination reported in the ST excerpts could just as easily be classified as self-contradictions. Given this, it seems advisable to avoid using the term *hallucination* and instead ask evaluators to focus on identifying self-contradiction. Notably, on her task feedback form, successful Student No. 20 observed that HT0 was the only text to mention specific places in a consistent way.

6.2.5 Unnecessary technical jargon

The four students who noted unnecessary technical jargon in HT0 all cited the same two examples: *grassazione*[9] and *rettifilo*[10]. These uncommon terms appear to be part of Alberto Moravia's idiolect, suggesting that this category is prone to producing false positives. The unreliability of this criterion for determining artificial-generated Italian text is one of the key findings of this study.

6.2.6 Other anomaly categories

According to the data in Table 2, none of the remaining categories proved effective in helping participants identify the ST excerpts.

6.3 Preparatory training

Since there was no initial control experiment conducted without preparatory training, it is hard to determine whether the training contributed to the success of the two students who performed well. Regardless, the fact that only 2 out of the 23 students (8.70%) were able to identify ST after training cannot be considered a successful outcome. Furthermore, as reported in Section 4.2, following the classroom discussion, half of the students deemed the training insufficient.

Successful Student No. 8 mentioned that, over the past year, she had often used GenAI tools (particularly ChatGPT) for reformulating, summarizing and occasionally translating texts, which are among the tasks some professional translators report they use GenAI for in their workflow (Farrell, 2025b). She suggested that this experience had helped her become familiar with the "distinctive writing style of GenAI". Taken together, these observations highlight the importance of providing training on how to use GenAI effectively for such tasks in translation courses.

6.4 Text excerpt length

Clark et al. (2021) truncated their text excerpts at the first end-of-sentence after reaching 100 words, while Dou et al. (2022) used whole paragraphs ranging from 80 to 145 tokens. In contrast, this experiment used sequential excerpts of between 268 and 467 words (Table 1). As mentioned in Section 4.2, none of the students found the excerpts too long, and after the discussion, the majority did not feel they needed to be any longer.

However, it seems plausible that low burstiness and self-contradiction would be easier to identify in longer excerpts (see also Section 6.2.2). In the case of short stories, these excerpts could potentially consist of the entire text.

6.5 Prompting to avoid anomalies

ST2 was generated using a prompt that specifically instructed ChatGPT-4o to avoid most of the tell-tale textual anomalies the students had been trained to identify. This seems to have been effective, since this story was rated, on average, as the most human-like of the seven analysed, with an overall mean score of 3.0. However, this prompting did not seem to successfully mislead Plagramme AI detector, even though ST2 received the second-lowest probability of being artificial among the six AI-generated stories (83%).

6.6 Need for synthetic-text editing

The need for translation students to be familiar with STE techniques stems from the hypothesis that demand for *traditional translation* is likely to decline, while demand for STE will probably grow.

The existence of this demand is in turn based on two assumptions: first, that readers are able to distinguish between ST and HT, and second, that they actually prefer reading HT.

Regarding the first assumption, as noted in Section 4.2, none of the students in this study

[8] Old-fashioned way of saying "I didn't stop for a moment".
[9] Armed robbery.
[10] Straight stretch of road.

considered distinguishing between ST and HT a pointless exercise, and the findings suggest that some individuals are indeed capable of doing so.

As for the second assumption, a study by Zhang and Gosline (2023) found that advertising content (in English) generated by GenAI, as well as human-created advertising content augmented by GenAI (i.e., automatically edited), was perceived as higher quality than content produced solely by human experts. Similarly, a study by Porter and Machery (2024) revealed that AI-generated poetry is indistinguishable from human-written verse and is rated more favourably.

Consequently, it would seem that STE may not be necessary for all genres of text.

7 Conclusion

The low number of students who were able to distinguish between the two kinds of text in this experiment, even after training, suggests that the guidance given needs redesigning. The examples highlighted by the participants indicate that the preparatory exercises should focus on identifying self-contradiction and assessing variability in syntactic structures and lexical distributions, known as burstiness.

Moreover, general training on the use of GenAI as a tool in the translation process, apart from being an essential part of any modern translation course, could also help students better identify ST.

Regarding the length of texts, it would be worthwhile experimenting with longer excerpts.

Lastly, further research should also explore whether readers genuinely prefer human-written text and whether STE, which seeks to make ST sound more human-like, is actually necessary at all.

Carbon impact statement

The study described in this paper involved seven queries made using ChatGPT-4o. According to a widely cited figure (Wong, 2024), each query generates approximately 4.32 g of CO_2 emissions. As a result, the entire experiment produced an estimated total of 30.24 g of CO_2, excluding the emissions generated from several hours of internet browsing for background research.

Acknowledgments

The research project reported in this paper has received funding from the International Center for Research on Collaborative Translation at the IULM University, Milan, Italy.

References

Chaka Chaka. 2023. *Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools*. Journal of Applied Learning & Teaching, July 2023. https://doi.org/10.37074/jalt.2023.6.2.12

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, Noah A. Smith. 2021. *All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text*. 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). https://aclanthology.org/2021.acl-long.565/

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, Yejin Choi. 2022. *Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text.* https://doi.org/10.48550/arXiv.2107.01294

Michael Farrell. 2025a. *Editing synthetic text from generative artificial intelligence: two exploratory case studies.* Proceedings of the 46th Conference Translating and the Computer, Luxembourg, November 18-20, 2024. https://www.tradulex.com/varia/TC46-luxembourg2024.pdf#page=35

Michael Farrell. 2025b. *Survey on the use of generative artificial intelligence by professional translators.* Proceedings of the 46th Conference Translating and the Computer, Luxembourg, November 18-20, 2024. https://www.tradulex.com/varia/TC46-luxembourg2024.pdf#page=23

Jonathan Gillham. 2024. *How To Identify AI-Generated Text?. Blog of Originality.ai AI & Plagiarism Detector.* https://originality.ai/blog/identify-ai-generated-text

Brian Porter and Edouard Machery. 2024. *AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably*. Sci Rep 14, 26133. https://doi.org/10.1038/s41598-024-76900-1

Vinnie Wong. 2024. *Gen AI's Environmental Ledger: A Closer Look at the Carbon Footprint of ChatGPT*. Piktochart.com. https://piktochart.com/blog/carbon-footprint-of-chatgpt/

Yunhao Zhang and Renée Gosline. 2023. *Human Favoritism, Not AI Aversion: People's Perceptions (and Bias) Toward Generative AI, Human Experts, and Human-GAI Collaboration in Persuasive*

Content Generation. Cambridge University Press. https://doi.org/10.1017/jdm.2023.37

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, Jimmy Ba. 2022. *Large language models are human-level prompt engineers.* Published as a conference paper at ICLR 2023. https://doi.org/10.48550/arXiv.2211.01910

Appendix A

The initial prompt generated by ChatGPT-4o was:

Generate a text in Italian that is approximately 1800 words long, including the title. The text should be a short story that explores themes of fear, courage, and moral dilemmas. It should feature a young protagonist who, after being influenced by a romantic interest, decides to write a threatening letter to the owner of a villa. The story should include vivid descriptions of the setting, the protagonist's thought process, the actual writing and delivery of the letter, and the psychological consequences that follow. The narrative should convey the protagonist's initial bravado, followed by increasing anxiety and fear as the reality of their actions sets in. The story should conclude with the protagonist retrieving the letter in a desperate attempt to avoid the consequences of their actions, only to be left questioning their courage and moral standing.

Appendix B

Replies to the debriefing questionnaire before and after the class discussion.

	Before	After
I agree	0	0
Maybe	1	4
I disagree	**6**	**3**
I don't know	1	1

Table 3: The task is pointless. Humans cannot distinguish between ST and HT, regardless of the training they receive.

	Before	After
I agree	4	4
Maybe	**3**	**1**
I disagree	0	0
I don't know	1	3

Table 4: The preparatory training was insufficient. Additional practice is needed to effectively identify textual anomalies.

	Before	After
I agree	0	**4**
Maybe	**2**	4
I disagree	4	0
I don't know	2	0

Table 5: Some of the textual anomalies we were searching for can also be found in HTs.

	Before	After
I agree	0	1
Maybe	1	2
I disagree	**6**	**3**
I don't know	1	2

Table 6: Searching for textual anomalies is not the right approach for this task. The results would have been better if we had relied on intuition.

	Before	After
I agree	1	0
Maybe	3	0
I disagree	3	**7**
I don't know	**1**	1

Table 7: The texts were too lengthy. The task would have been easier if shorter excerpts had been provided.

	Before	After
I agree	0	**2**
Maybe	**1**	1
I disagree	6	5
I don't know	1	0

Table 8: The texts were too brief. The task would have been easier if we had been given the entire short story.

MT or not MT? Do translation specialists know a machine-translated text when they see one?

Rudy Loock, Nathalie Moulard, Quentin Pacinella
Université de Lille / rudy.loock@univ-lille.fr, nathalie.moulard@univ-lille.fr, quentin.pacinella@univ-lille.fr

Abstract

In this article, we investigate translation specialists' capacity to identify raw machine translation (MT) output in comparison with so-called "human" translations produced without any use of MT. Specifically, we measure this capacity via an online activity, based on different criteria: (i) degree of expertise (translation students vs. professionals with at least 5 years' experience), (ii) MT engine (DeepL, Google Translate, Reverso, ChatGPT), and (iii) length of input (1-3 sentences). A complementary, qualitative analysis, based on participants' feedback, provides interesting insight on how they discriminate between raw MT output and human translations.

1 Introduction

With the advent of neural machine translation (NMT) in the middle of the 2010s, which allowed for a surge in the quality of MT output in comparison with previous systems (e.g. Toral & Sánchez-Cartagena 2017, Van Brussel et al. 2018), there has been a lot of discussion on the comparison between machine-translated texts and so-called human translations. Some studies have shown that errors in NMT output resemble errors to be found in translations produced by human professionals, making them harder to detect and less transparent for both professionals and translation trainees (e.g. Castilho et al. 2017a/2017b, Yamada 2019). While some research has gone so far as to claim "human parity" (Hassan et al. 2018), it has been shown that machine-translated texts do show specific linguistic properties that distinguish them from human-produced translations. Among such features are for example a lesser lexical variety, syntactic normalization, or terminological inconsistencies, all of which make so-called "machine-translationese" a reality (e.g. Vanmassenhove et al. 2019, Vanmassenhove et al. 2021; Loock 2020; De Clercq et al. 2021).

As a consequence, the traditional distinction between two types of translation corrections have been maintained: revision, which is the correction of human-produced translations, and post-editing, which is the correction of machine-translated texts. The industry has so far maintained this distinction, with the existence of two ISO standards, sometimes including a revision step after MT post-editing. Translation programs have set up distinct classes to teach both tasks, even separate models to evaluate the two competences (Kontinnen et al. 2021).

There is however evidence that the line between the two might be blurring (Daems & Macken 2021, Do Carmo & Moorkens 2021), all the more so as the line between human-translated and machine-translated texts is blurring: even outside MTPE projects, professionals use MT output as a source of inspiration.

If one wants to maintain the distinction between the two tasks, leading to different types of corrections, then this means that professional revisers and post-editors should know the origin of the translations that they are supposed to correct. And as this is not always the case – with even some cases where revisers are asked to *revise* a machine-translated text without being properly informed[1] – it seems important to evaluate professionals' capacity to discriminate between raw MT output and translations produced by professional

[1] Unfortunately, there are no surveys or studies about this phenomenon, but there are many testimonies from freelance translators working with translation agencies saying that this does regularly happen these days.

Proceedings of Machine Translation Summit XX Volume 1, pages 442–454, June 23-27, 2025

translators, henceforth human translations (HTs). Not only is this a technical competence, but sociological considerations also need to be taken into account (Daems & Macken 2021), as translators tend to mistrust MT output more than human-translated texts, which can lead to "over-editing" (Nitzke & Gros 2021) in the case of MTPE, although the picture is more complicated than that, as shown by Daems & Macken (2021).

2 Research question

Therefore, in this paper, our aim is to check whether it is possible for translation experts to identify raw MT outputs (i.e. without any post-editing) among HTs, with a focus on 2 types of users: students in their final year of a master's training program, right before they join the translation industry, and (ii) translation professionals with at least 5 years' professional activity. Our aim is to investigate whether experience, the MT engine, the length of the input, but also the original text itself has an influence on users' ability to discriminate between MT and HT.

3 Methodology

In order to answer our research question, we set up an online exercise where participants were shown 4 translations into French of English sentences for a series of 20 items ranging from 1 to 3 sentences. The EN-FR translations consisted of a mix between raw MT outputs obtained through 4 different tools (see below) and HTs produced by experienced professional translators. For each of the 20 items, there were between 0 and 4 raw MT outputs, the rest being HTs. A total of 221 participants were recruited, students enrolled in their second and final year of a master's translation program (MA2) in France and translation professionals with at least 5 years' activity. All of them had French as their native language. Below we provide detailed information on the data, the participants, and the exercise.

3.1 Data

The data used in our online exercise comes from 3 main sources. First, 2 articles were selected from the US website of National Geographic, one on tardigrades[2] and the other on Ozempic, a weight loss drug[3]. These texts were chosen as they both deal with a specialized topic and belong to a specific register (scientific press), presenting both terminological and stylistic issues for translation. They were both published in the summer of 2024 and no translation on the French website was available when the experiment was conducted.

Second, the 2 articles were translated with 4 different tools, 3 now traditional online translators (DeepL, Google Translate, Reverso Translation)[4] and ChatGPT v. 4o, a generative AI tool not specifically developed for translation but capable of achieving translation tasks[5]. All outputs were retrieved in October 2024.

Second, 8 professional translators were recruited so that each article could be submitted to 4 different professionals with the instruction to provide a natural, professional-sounding translation into French. They were specifically asked not to use any MT of any sort, but they were free to use any other tools they wanted.

For our experiment, we did not use all of the 2 texts, but the first 737 words for the 1st text and the first 836 for the 2nd text. The number of words retained in our experiment for each text is provided in Table 1.

Type of Text	Text 1	Text 2
Original text (EN)	737	836
HT1	832	1218
HT2	814	1095
HT3	924	1058
HT4	1023	1079
MT1 (DeepL)	839	1065
MT2 (Google Translate)	840	1105
MT3 (Reverso)	824	1028
MT4 (ChatGPT-4o)	822	988

Table 1: Number of words for each text in our data set

Each original text was broken into 20 items, containing from 1 to 3 sentences, and aligned with the 8 translations (4 HT and 4 MT). For each item,

[2] https://www.nationalgeographic.com/science/article/water-bear-tardigrade-fossil-amber-evolution
[3] https://www.nationalgeographic.com/science/article/ozempic-mounjaro-lower-risk-10-cancers-chronic-disease
[4] Because of the limits in the number of characters (5,000 for DeepL and Google Translate, 2,000 for Reverso), the texts had to be split but we were careful to always include the beginning of the texts to ensure the inputs remained coherent.
[5] The prompt used in ChatGPT was a zero-shot prompt (*Translate the following into French*).

only 4 translations were retained for submission to the respondents, with a random selection, the only constraint being that each text, whether translated by a professional or machine-generated, should be used the same number of times (20). Items contained from 0 to 4 MTs and similarly from 0 to 4 HTs. For each text, 12 items consisted of 1 sentence, 4 of 2 sentences, and 4 of 3 sentences. The order in which the 4 translations were presented was also random. However, the distribution was similar for the 2 texts, with a total of 40 MTs and 40 HTs in total.

The reason why we used 2 texts was to check whether the text had an influence on the results.

3.2 Participants

We submitted our online exercise via different Google Forms (responses were anonymous and no personal data such as-emails were collected) to 2 different types of respondents:

(i) Students enrolled in their second and final year of an MA program (MA2) in a French university (n=187). All of them were native speakers of French. They received the link to the exercise through one of the teachers in their program via the AFFUMT association (French association of translation training programs). They neither received credits nor compensation for fulfilling the task, and they were free to do it or not, either in class or at home. A question prior to the exercise revealed that 50.8% of students had received MT training and 71.1% training on MT post-editing.

(ii) Translation professionals (n=34), all of them native speakers of French. As one of our goals was to check whether expertise/experience had an impact on users' capacity to identify MT output, they were required to have at least 5 years' experience (between 5 and 10 years for 11 of them, between 10 and 15 years for 7 of them, and more than 15 years for 16 of them). They were all contacted by e-mail and received no compensation for their participation. A question prior to the exercise revealed that 44.1% of them had received MT training and 55.9% training on MT post-editing, while 82.3% of them had already done post-editing tasks.

3.3 Online exercises

In total, 2 online exercises were prepared, 1 for each text, to be submitted to the 2 categories of respondents. The items were similar and presented in the same way and order.

Each respondent was submitted to 20 questions, that is 20 inputs in English (ranging from 1 to 3 sentences) and 4 different translations (respondents were systematically shown the English original input alongside the 4 translations). They were asked to tick the boxes next to the translations which they thought were raw MT outputs. In appendix we provide the 20 items for the online exercise with text 1 (tardigrades). It was clearly specified in the instructions that the MT outputs were raw, without any post-editing at all, and that among the 4 translations, there could be from 0 to 4 MTs, the other translations being produced by translation professionals. All along the exercise, they were never told whether their answers were correct or not.

Participants were also asked before starting the exercise whether they felt confident in identifying MT output. At the end of the exercise, they were asked how difficult they had found the task. They also had the opportunity if they wanted to provide verbatim feedback on what helped them identify MT outputs.

Only the introductory questions differed between the 2 groups. Students were asked to confirm they were enrolled in an MA2 from a translation program and native speakers of French. Professionals were asked to confirm that they had at least 5 years' experience and were native speakers of French.

4 Results

4.1 General results

Our general results show that respondents got an average score of 5.68 out of 20, with 5.53 for students and 6.52 for professionals, and results ranging from 0 to 13 out of 20. Figure 1 provides a summary of the results.

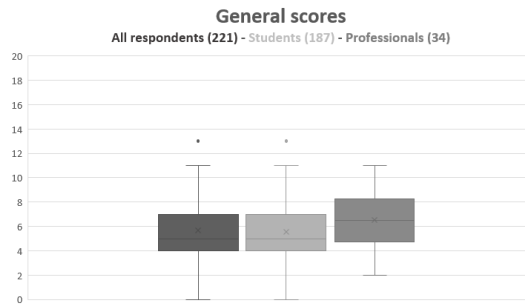


Figure 1: General scores (out of 20 points)

This might seem a poor result at first sight, but it is important to remember that in order to get the

point, respondents had to correctly identify the origin of the 4 different translations. When one focuses instead on the correct identification rate of MTs and HTs in general, results are actually much better: respondents identified raw MT outputs correctly in 65.48% of cases (65.35% for students and 66.18% for professionals) and identified HTs correctly in 76.59% of cases (76.01% for students and 79.79% for professionals). The results, shown in Figure 2, mean that respondents have a tendency to misidentify MTs as HTs more often than they misidentify HTs as MTs. A chi-square test was conducted to compare the success rates of the 2 groups, students and professionals. The results show a significant difference between the 2 groups for HT identification ($p<.001$) but not for MT identification ($p=.396$), while overall results show a significant difference ($p<.001$).

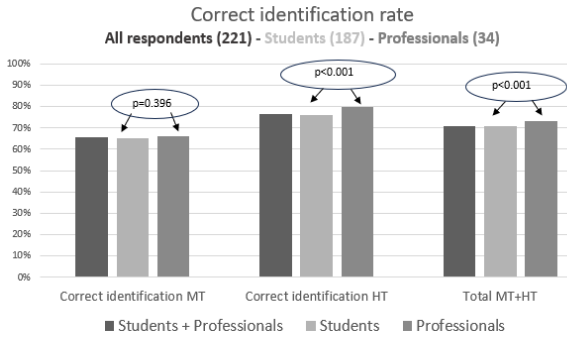


Figure 2: Identification rates (in%)

If one compares results depending on the text used for the exercise (see Figure 3), differences can only be spotted for professionals, who found it more difficult to identify MTs for the second text.

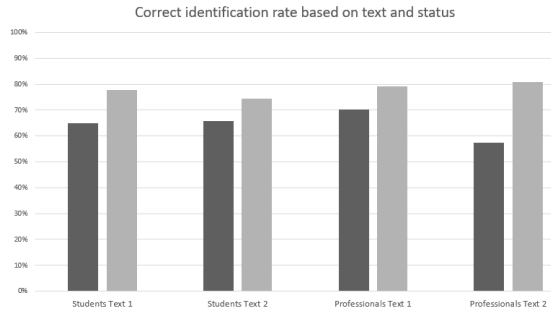


Figure 3: Identification rates students vs. professionals (in%)

4.2 Results depending on MT tool

If one compares the correct identification rate depending on the MT tool used to generate the MT outputs (DeepL, Google Translate, Reverso, ChatGPT4), the results show that the rate is the lowest for DeepL, followed by Google Translate and ChatGPT4, and then by Reverso (see Figure 4).

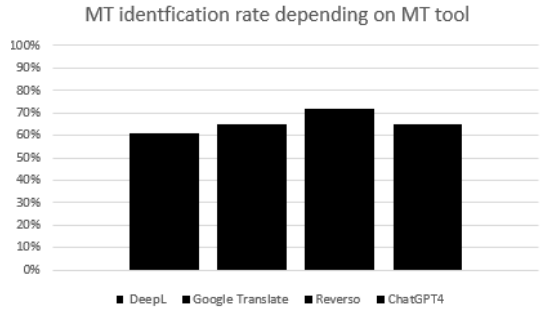


Figure 4: Identification rates (in%) depending on MT tool

This means that among the 4 MT engines, DeepL is the one that produced raw outputs that more often passed as HTs for our respondents. Statistically, only the difference between Google Translate and ChatGPT4 is not significant ($p>0.05$), which leads to the following result in terms of performance for the 4 tools under investigation:

DeepL > Google Translate = ChatGPT4 > Reverso[6]

A comparison between results for the 2 texts (see Figure 5) shows some difference, though: while

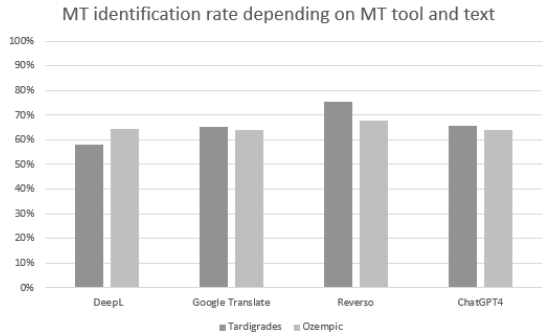


Figure 5: Identification rates (in%) depending on MT tool and text

[6] This reads as follows: DeepL provided better results (i.e. its outputs were more frequently identified as HT) than Google Translate and ChatGPT4 which showed similar results while themselves providing better results than Reverso.

results for text 1 (tardigrades) are similar to the general trend, for text 2 (weight loss drug) differences between the 4 engines are not statistically significant.
This leads to the following result in terms of performance for the 4 tools under investigation:

Text 1 (Tardigrades):
DeepL > Google Translate = ChatGPT4 > Reverso

Text 2 (Ozempic):
DeepL = Google Translate = ChatGPT4 = Reverso

Finally, if one compares results for students and professionals, there are few differences (see Figure 6): professionals seem to confuse DeepL MT outputs with HTs more often than students (correct identification rate of 58.55% vs. 61.71%), but they identify Reverso and Google Translate MT outputs more easily (correct identification rates of 77.85% and 66.48% vs. 70.54% and 64.45% respectively).

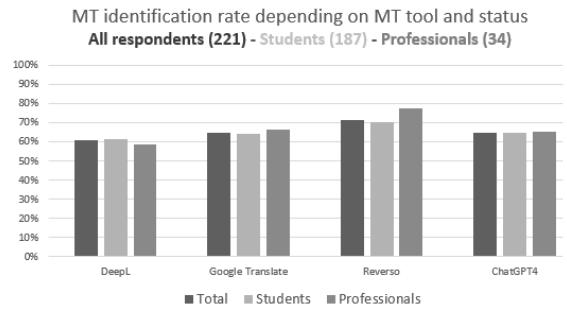


Figure 6: Identification rates (in%) depending on MT tool and status

4.3 Results for HTs

The results for the identification of the 8 HTs produced by 8 different translation professionals showed much more variation, with correct identification rates ranging from 51.98% to 80.55% (Figure 7). This might be due to differences in quality (see verbatim comments in section 4.5) and would require further investigation.

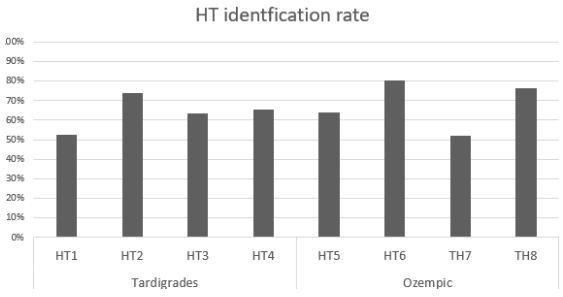


Figure 7: Identification rates (in%) for human translations

4.4 Results depending on length of input

One of our hypotheses when designing the experiment was that the longer the input, the better the identification of the origin of the translation would be, as MT is known for encountering difficulties to deal with the way sentences connect between each other. Such a hypothesis is not validated by our results shown in Figure 8: the comparison between results for short inputs (8-20 words), average-length inputs (20-40 words), and longer inputs (40-85 words) does not reveal a systematic pattern. For example, while results do improve for students with Text 2, they actually deteriorate with Text 1. Our results are here inconclusive.

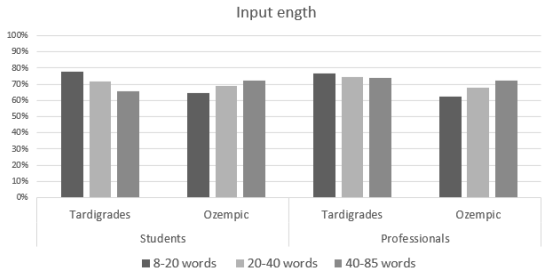


Figure 8: Identification rates (in%) depending on length of input

4.5 Respondents' perception before/after the exercise and feedback

Before starting the activity, respondents were asked whether they considered themselves capable of identifying raw MT outputs among professional translations. Among students, 59% fully or rather agreed with the assertion that they were capable of achieving such a task, a proportion that rose to 74% among professionals. However, when asked after the exercise whether they had found it easy or difficult, 67% of students and 53% of professionals found it difficult or very difficult, with only 3% of students and 15% of professionals finding the exercise to be easy (no respondent found it very easy).

We also gave our respondents the opportunity to provide verbatim comments on how they were able to identify raw MT outputs. We did not ask for feedback for each individual item for fear of survey

fatigue, but asked for some general feedback at the end of the questionnaire with an optional question. Quite a number of respondents did provide such feedback (n=111/221), which revealed what kind of elements according to them helped them discriminate between raw MTs and HTs.

Many mentioned that when translations were literal, they considered them to be MTs rather than HTs, both for lexical choices (e.g. literal translation of the verb *say* by *dire,* literal translations leading to repetitions or atypical collocation phenomena) and syntactic choices (e.g. same word order or syntactic constructions). For instance, the translation of the verb *say* by its French direct equivalent *dire* in sentence (1) while this verb is hardly ever used in the press genre, combined with a literal translation of *noting* with *notant*, quite unnatural in French, seems to have led to the identification of MT: 84.4% of students and 87% of professionals identified (1a) as MT. However, the use of the verb *indiquer* ('to indicate') and the gerund *en précisant* ('by specifying') in (1b) led to a correct identification of HTs by 88.9% of students and 91.3% of professionals.

(1) Just a few paleontologists study fossil tardigrades, Mapalo says, noting that some colleagues react with surprise that any fossil tardigrades are known at all.
a. Seuls quelques paléontologues étudient les fossiles de tardigrades, dit Mapalo, notant que certains de ses collègues sont surpris d'apprendre que des fossiles de tardigrades existent même.
b. « Seuls quelques paléontologues étudient les fossiles de tardigrades », indique Marc Mapalo, en précisant que « certains de ses collègues sont même surpris que des fossiles de tardigrades puissent exister ».

Similarly, the association of the verb *endurer* and *certaines des conditions les plus difficiles,* a literal translation of *endured some of the harshest conditions*, seems to have led 95.6% of students and 91.3% of professionals to correctly identify the output as MT, as opposed to *supporter des conditions difficiles*.

A calque of the word order in (2a) has led respondents to identify such a translation as an MT output by 85.6% of students and 91.3% of professionals, while a reordering with a translation beginning with *en comprenant quand* ('by understanding when') has led to only 15.6% of students and 4.3% of professionals considering the HT in (2b) to be MT (note that other differences such as the nominalization strategy to translate *how* and *why* may also have played a role).

(2) "Knowing when cryptobiosis evolved in tardigrades can help us contextualize how and why they gained this mechanism," Mapalo says. Tardigrades likely evolved in the seas before spreading onto land, he notes.
a. « Savoir quand la cryptobiose a évolué dans les tardigrades peut nous aider à contextualiser comment et pourquoi ils ont acquis ce mécanisme », explique Mapalo.
b. « En comprenant quand les tardigrades ont développé la cryptobiose, nous pouvons formuler des hypothèses sur la manière et la raison de l'apparition de ce mécanisme », explique Marc Mapalo.

Respondents also mentioned in their comments that reproducing the same word order for long sentences was a clear sign of MT. For instance, the literal translation in the MT output (3a) led to 93.8% of students and 90.9% of professionals identifying it as MT. On the other hand, the HTs starting with *dans le cadre d'une étude…* ('within the framework of a study') or *dans un article publié…* ('in an article published'), and showing a word order reorganization were identified as MT outputs only by 19.6%/6.2% of students and 18.2%/0% of professionals respectively.

(3) "The cardioprotective effect of semaglutide observed in people with obesity developed within months of drug initiation, well before meaningful weight loss had been achieved in most trial participants" in one 2022 trial, Daniel Drucker, a physician-scientist at the Lunenfield-Tanenbaum Research Institute at Mt. Sinai Hospital in Toronto, states in a commentary published Thursday in Science.
a. « L'effet cardioprotecteur du sémaglutide observé chez les personnes obèses s'est développé dans les mois suivant le début du traitement, bien avant qu'une perte de poids significative n'ait été obtenue chez la plupart des participants à l'essai » dans un essai de 2022, déclare Daniel Drucker, médecin-chercheur à l'Institut de recherche Lunenfield-Tanenbaum de l'hôpital Mt. Sinai à Toronto, dans un commentaire publié jeudi dans Science.

Finally, terminological errors, e.g. the use of *tartariens* ('tartarians') to translate *tardigrade folks*, or repetitions due to literal translations (less acceptable in French than in English), as well as inconsistencies in the use of punctuation were also mentioned as factors leading to MT identification.

On the other hand, explicitations, e.g. adding *le magazine* in front of *Science*, or word order reorganization as for example (3) were for the respondents signs that the translations were produced by a human.

What all of these comments reveal is that respondents searched for translation problems, which they automatically attributed to the fact that translations were generated by an MT engine. Only 1 respondent mentioned the fact that they wondered whether translation errors were due to MT or poor HTs, and 2 respondents mentioned that the HTs were not always high quality.

5 Discussion and conclusion

What the results of our experimentation show is that translation experts, whether professionals or students in their final year of studies in a master's program right before joining the translation industry, are capable of discriminating between raw MT output and professional "human" translations in 2 thirds of cases on average. Professionals perform better than students, but only slightly. This might be due to the fact that nowadays, most training programs include training on MT and post-editing, while not all professional translators have received such training. Among the 4 generic tools under investigation here, DeepL is the one that seems to provide the best outputs, since they are more frequently confused with human translations. Google Translate and ChatGPT follow, while Reverso provides raw outputs that are the most easily identifiable by our respondents. Our results also show that the choice of text may have an influence on the result, while the length of the input does not seem to. All of these results lead to the conclusion that machine translation should not be considered as ONE unique tool, but that the quality of any MT output depends on a number of factors, in particular the MT engine that is used and the text that is translated.

Our results also confirm the existence of "machine-translationese" (see introduction), since raw MT outputs show features, both lexical and syntactic, that can help differentiate them from texts produced by humans. This means that an automatic detection of MT outputs via a specific tool is a possibility that could be considered, although beyond the scope of this paper.

In terms of MT-related competences, our results show that in order to develop a good and relevant MT literacy as defined by Bowker & Buitrago Ciro (2019), it is important not to overestimate one's capacity to identify MT output, especially as the verbatim results show that respondents consider any translation error to be due to an MT engine rather than to a human being. Such a bias could lead to over-editing, a risk that is widely assumed in the case of post-editing for sociological reasons, although Daems & Macken (2021)'s experiment actually could not confirm it (their respondents brought more changes to MT outputs when they thought that they were actually revising human translations). However, it is also important not to underestimate one's competences: after all, our results reveal that respondents can identify MT outputs among HTs in 2 thirds of cases.

There are naturally some limitations to our study. The very first one is that our experiment deals with one language pair for one translation direction only (EN-FR) as well as one text type, and therefore the results cannot be generalized. It also needs to be acknowledged that as results provided by MT engines change over time, it is not possible to reproduce our experiment and obtain the same results as those obtained in October 2024. Second, we have used free, generic online translators, although professionals in the translation industry often use custom MT engines or professional paid versions of MT tools. It would be interesting to reproduce the same kind of experimentation with MT outputs from such tools. Third, we have not compared results from respondents who have received MT and/or PE training and those who have not. Finally, it would be relevant to conduct the same study with people who are not translation experts but rather experts in the fields related to the topics of the texts (zoology and medicine), and see whether, as experts of the terminology in these fields, they are better than translation specialists at identifying MT outputs. These aspects are left open for future research.

Acknowledgments

We would like to thank the 221 respondents who took part in our experiment on a voluntary basis. Special thanks go to the colleagues in the French translation programs who circulated the exercise among their MA2 students.

References

Lynne Bowker and Jairo Buitrago Ciro. 2019. *Machine Translation and global research: Towards improved machine translation literacy in the scholarly community*. Emerald Publishing, Bingley.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017a. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108: 109-120. https://doi.org/10.1515/pralin-2017-0013.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli-Barone, and Maria Gialama. 2017b. A comparative quality evaluation of PBSMT and NMT using professional translators. In *Proceedings of the Machine Translation Summit XVI*, v. 1, pages 116–131. https://aclanthology.org/2017.mtsummit-papers.10/.

Joke Daems and Lieve Macken 2021. Post-editing human translations and revising machine translations: impact on efficiency and quality. In Maarit Koponen, Brian Mossop, Isabelle S. Robert and Giovanna Scocchera (eds), *Translation Revision and Post-editing. Industry Practices and Cognitive Processes*. Routledge, London/New York, pages 50–70. http://dx.doi.org/10.4324/9781003096962-5.

Orphée De Clercq, Gert de Sutter, Rudy Loock, Bert Cappelle, and Koen Plevoets. 2021. Uncovering machine translationese using corpus analysis techniques to distinguish between original and machine-translated French. *Translation Quarterly*, 101:21-45.

Félix do Carmo and Joss Moorkens. 2021. Differentiating editing, post-editing and revision. In Maarit Koponen, Brian Mossop, Isabelle S. Robert and Giovanna Scocchera (eds), *Translation Revision and Post-editing. Industry Practices and Cognitive Processes*. Routledge, London/New York, pages 35–49.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. https://doi.org/10.48550/arXiv.1803.05567.

Maarit Konttinen, Brian Mossop, Isabelle S. Robert, and Giovanna Schocchera (eds). 2021. *Translation Revision and Post-editing: Industry Practices and Cognitive Processes*. Routledge, London/New York.

Rudy Loock. 2020. No more rage against the machine: How the corpus-based identification of machine-translationese can lead to student empowerment. *The Journal of Specialised Translation*, 34:150-170.

Jean Nitzke and Anne-Kathrin Gros. 2021. Preferential changes in revision and post-editing. In Maarit Koponen, Brian Mossop, Isabelle S. Robert and Giovanna Scocchera (eds), *Translation Revision and Post-editing. Industry Practices and Cognitive Processes*. Routledge, London/New York, pages 21–34.

Antonio Toral Ruiz and Victor M. Sanchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1701.02901

Laura Van Brussel, Arda Tezcan, and Lieve Macken. 2018. A fine-grained error analysis of NMT, SMT and RBMT output for English-to-Dutch. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* pages 3799–804. https://aclanthology.org/L18-1600/

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII v. 1: Research Track*, pages 222–232. https://www.aclweb.org/anthology/W19-6622/

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, pages 2203–2213. 10.18653/v1/2021.eacl-main.188.

Masuru Yamada. 2019. The impact of Google Neural Machine Translation on post-editing by student translators. *The Journal of Specialised Translation*, 31, 87-106.

Appendix | The 20 items extracted from Text 1 (1-20) alongside their 4 translations submitted to respondents (a-d)

1. They survived an apocalypse—by sleeping through it (Note : il s'agit du titre de l'article)
a. Ils ont survécu à l'apocalypse en dormant.
b. Ils ont survécu à une apocalypse – en dormant
c. Des organismes survivent à une apocalypse en restant endormis
d. Pour survivre à l'apocalypse, ils ont fait le choix de dormir

2. The specimens provide insight into how tardigrades evolved cryptobiosis, a temporary and almost complete shutdown of bodily processes.
a. Grâce à l'étude de certains spécimens, les scientifiques en ont appris davantage sur la manière dont les tardigrades ont pu entrer en cryptobiose, une extinction temporaire quasiment complète des processus corporels.
b. Les échantillons ci-dessous nous aident à comprendre ce qui a déclenché le développement de la cryptobiose, un arrêt temporaire et quasi total de l'organisme, chez les tardigrades.
c. Les échantillons nous fournissent des informations sur la façon dont les tardigrades ont développé la capacité de cryptobiose, un arrêt temporaire et presque complet de leur métabolisme.
d. Les spécimens permettent de comprendre comment les tardigrades ont évolué vers la cryptobiose, un arrêt temporaire et presque complet des processus corporels.

3. Tardigrades are survivors. For more than 500 million years, the microscopic "water bears" have spread all over the planet and endured some of the harshest conditions Earth has to offer.
a. Les tardigrades sont des survivants. Depuis plus de 500 millions d'années, les « ours d'eau » microscopiques se sont répandus sur toute la planète et ont enduré certaines des conditions les plus difficiles que la Terre peut offrir.
b. Les tardigrades sont des survivants. Depuis plus de 500 millions d'années, les microscopiques « oursons d'eau » se sont répandus partout sur la planète et ont enduré certaines des conditions les plus extrêmes que la Terre ait à offrir.
c. Les tardigrades sont de véritables survivants. Depuis plus de 500 millions d'années, ces « oursons d'eau » se sont répandus sur la planète et ont traversé certains des environnements les plus difficiles connus sur Terre.
d. Les tardigrades sont des survivants. Pendant plus de 500 millions d'années, ces « oursons d'eau » microscopiques se sont propagés sur toute la planète et ont supporté les conditions les plus difficiles que la Terre peut offrir.

4. Now a new analysis of ancient tardigrades in a piece of Cretaceous amber has not only clarified the timeline of tardigrade evolution, but hints how the tiny animals have been able to survive disasters that drove other forms of life to extinction.
a. Une nouvelle analyse d'anciens tardigrades dans un morceau d'ambre du Crétacé a permis non seulement de clarifier la chronologie de l'évolution des tardigrades, mais aussi de comprendre comment ces petits animaux ont pu survivre à des catastrophes qui ont conduit d'autres formes de vie à l'extinction
b. Aujourd'hui, une nouvelle analyse d'anciens tardigrades dans un morceau d'ambre du Crétacé a non seulement clarifié la chronologie de l'évolution des tardigrades, mais a également permis de comprendre comment ces minuscules animaux ont pu survivre à des catastrophes qui ont conduit d'autres formes de vie à l'extinction
c. Une nouvelle analyse des tardigrades anciens dans un morceau d'ambre du Crétacé a non seulement clarifié la chronologie de l'évolution des tardigrades, mais suggère également que les minuscules animaux ont pu survivre aux catastrophes qui ont conduit à l'extinction d'autres formes de vie.
d. Maintenant, une nouvelle analyse d'anciens tardigrades emprisonnés dans un morceau d'ambre du Crétacé a non seulement clarifié la chronologie de l'évolution des tardigrades, mais laisse entendre comment ces minuscules animaux ont pu survivre à des catastrophes qui ont entraîné l'extinction d'autres formes de vie.

5. The tiny critters were trapped in tree sap in prehistoric Canada between 83 and 72 million years ago, when giant tyrannosaurs and horned dinosaurs roamed the same conifer forests.
a. Les oursons d'eau ont été piégés dans de la sève d'arbre au Canada entre 83 et 72 millions d'années avant notre ère, lorsque de gigantesques tyrannosaures et des dinosaures cornus erraient dans les mêmes forêts de conifères.
b. Ces minuscules créatures ont été piégées dans de la sève d'arbre dans le Canada préhistorique, entre 83 et 72 millions d'années, à l'époque où des tyrannosaures géants et des dinosaures à cornes parcouraient les mêmes forêts de conifères.
c. Les petites bestioles ont été retrouvées piégées dans de la sève d'arbre datant du Canada préhistorique, soit il y a entre 83 et 72 millions d'années, à une époque où les géants tyrannosaures et dinosaures à cornes parcouraient encore ces mêmes étendues de conifères.
d. Ces minuscules créatures, qui côtoyaient d'immenses tyrannosaures et tricératops dans les forêts de conifères, sont restées prisonnières de la sève de ces arbres au Canada préhistorique, il y a 72 à 83 millions d'années.

6. One of the tardigrades is a species paleontologists have seen before. Named Beorn leggi, the tardigrade was the first fossil species ever discovered by paleontologists. But Harvard University paleontologist Marc Mapalo and his colleagues also found a second, never-before-seen species, Aerobius dactylus.

a. L’un des tardigrades est une espèce que les paléontologues ont déjà vue. Le tardigrade, appelé Beorn leggi, est la première espèce fossile jamais découverte par les paléontologues. Mais le paléontologue Marc Mapalo de l’Université Harvard et ses collègues ont également découvert une deuxième espèce, jamais observée auparavant, Aerobius dactylus.
b. L’une des espèces de tardigrades identifiées est bien connue des paléontologues. De son joli nom Beorn leggi, ce tardigrade a été la première espèce fossilisée jamais découverte par les paléontologues. Toutefois, Marc Mapalo, paléontologue à l’Université d’Harvard, et ses collègues ont également découvert une seconde espèce jamais vue auparavant : Aerobius dactylus.
c. L’un d’eux est déjà bien connu des paléontologues : Beorn leggi, la première espèce fossile jamais découverte. Cependant, l’équipe du paléontologue Marc Mapalo, de l’Université d’Harvard, a découvert une seconde espèce jusqu’alors inconnue, qu’elle a appelée Aerobius dactylus.
d. L'espèce de tardigrades Beorn leggi a déjà été observée par les paléontologues. Il s'agit en effet de la première espèce fossile qu'ils ont découverte. En revanche, le paléontologue de l'université Harvard Marc Mapalo et ses collègues ont découvert une deuxième espèce jamais observée auparavant, l'Aerobius dactylus.

7. The researchers named the new species and used it and the handful of other ancient species known to science to analyze the evolutionary history of tardigrades in Communications Biology earlier this month.

a. Les chercheurs ont nommé cette nouvelle espèce et l'ont utilisée, avec quelques autres espèces anciennes connues de la science, pour analyser l'histoire évolutive des tardigrades dans un article publié plus tôt ce mois-ci dans Communications Biology.
b. Les chercheurs ont nommé la nouvelle espèce et l'ont utilisée, ainsi que la poignée d'autres espèces anciennes connues de la science, pour analyser l'histoire évolutive des tardigrades dans la revue Communications Biology, publiée au début du mois.
c. Après lui avoir attribué un nom, les chercheurs se sont servis de cette nouvelle espèce et de la poignée d’autres espèces anciennes déjà connues pour analyser l’histoire de l’évolution des tardigrades et ont publié leurs conclusions dans Communications Biology plus tôt ce mois-ci.
d. Les chercheurs ont donc donné son nom à la nouvelle espèce, puis l’ont utilisée ainsi qu’une poignée d’autres espèces préhistoriques connues de la science afin d’analyser la chronologie de l’évolution des tardigrades. Cette analyse a été publiée plutôt ce mois-ci, sur le site Communications Biology.

8. Fossilized within the ancient tree resin that forms today’s amber, the two tardigrades had been waiting decades for a good look. Paleontologists could barely make out the B. leggi fossil in the Canadian specimen when they first described it 1964. Now, thanks to enhanced imaging technology, Mapalo and colleagues were able to get a much more detailed look.

a. Fossilisés dans l’ancienne résine d’arbre qui forme aujourd’hui l’ambre, les deux tardigrades attendaient depuis des décennies un bon regard. Les paléontologues ont à peine pu distinguer le fossile de B. leggi dans le spécimen canadien lorsqu’ils l’ont décrit pour la première fois en 1964. Grâce à la technologie d’imagerie améliorée, Mapalo et ses collègues ont pu obtenir un regard beaucoup plus détaillé.
b. Fossilisés dans l'ancienne résine d'arbre qui forme l'ambre d'aujourd'hui, les deux tardigrades attendaient depuis des décennies de pouvoir être observés. Les paléontologues pouvaient à peine distinguer le fossile de B. leggi dans le spécimen canadien lorsqu'ils l'ont décrit pour la première fois en 1964. Aujourd'hui, grâce à une technologie d'imagerie améliorée, Mapalo et ses collègues ont pu obtenir un aperçu beaucoup plus détaillé.
c. Fossilisés dans de l’ancienne résine d’arbre devenue aujourd’hui de l’ambre, les deux tardigrades ont attendu des dizaines d’années avant de pouvoir être étudiés en détail. Lorsqu’ils ont décrit pour la première fois le spécimen canadien en 1964, les paléontologues pouvaient à peine distinguer le fossile de Beorn leggi. Aujourd’hui, grâce aux progrès de la technologie d’imagerie, Marc Mapalo et ses collègues ont pu profiter d’une vue bien plus détaillée.
d. Fossilisés dans l’ancienne résine d’arbre qui forme aujourd’hui l’ambre, les deux tardigrades attendaient depuis des décennies d’être examinés de plus près. Les paléontologues pouvaient à peine discerner le fossile de B. leggi dans le spécimen canadien lorsqu'ils l'ont décrit pour la première fois en 1964. Aujourd'hui, grâce à une technologie d'imagerie améliorée, Mapalo et ses collègues ont pu l’examiner en détail.

9. “Lots of tardigrade folks have pondered these fossils over the last 60 years but there was a hard limit to how much could be gleaned because the tardigrades were really small and a bit obscured by the amber,” says New Jersey Institute of Technology biologist Phil Barden, who was not involved in the new study. The animals are so small, he notes, that the tiny claws on their feet are about one tenth the width of a human hair.

a. « Beaucoup de tartariens ont réfléchi à ces fossiles au cours des 60 dernières années, mais il y avait une limite stricte à la quantité qu’ils pouvaient récolter parce que les tardigrades étaient vraiment petits et un peu obscurcis par l’ambre », dit le biologiste du New Jersey Institute of Technology, Phil Barden. qui n’a pas participé à la nouvelle étude. Les animaux sont si petits, note-t-il, que les petites griffes sur leurs pieds font environ un dixième de la largeur d’un poil humain.

b. D’après Phil Barden, biologiste au New Jersey Institute of Technolgy, qui n’a pas participé à la nouvelle étude, « De nombreux spécialistes des tardigrades se sont penchés sur ces fossiles au cours des 60 dernières années, mais la quantité d’informations à recueillir était très faible, car les tardigrades étaient vraiment minuscules et un peu occultés par l’ambre. Ces animaux sont tellement petits, ajoute-t-il, que les minuscules griffes au bout de leurs pattes sont environ dix fois moins épaisses qu’un cheveu humain ».
c. « Au cours des 60 dernières années, de nombreux spécialistes des tardigrades ont étudié ces fossiles, mais ils n’étaient pas en mesure d’en extraire beaucoup d’informations en raison de la taille très réduite des spécimens et de l’obscurcissement provoqué par l’ambre », explique le biologiste Phil Barden, du New Jersey Institute of Technology, qui n’a pas participé à la nouvelle étude. Il ajoute que ces animaux sont si petits que leurs minuscules griffes font environ un dixième de la largeur d’un cheveu humain.
d. « De nombreux chercheurs de tardigrades ont étudié ces fossiles au cours des 60 dernières années, mais il y avait une limite stricte à ce que l’on pouvait en glaner, car les tardigrades étaient vraiment petits et un peu cachés par l’ambre », explique Phil Barden, biologiste au New Jersey Institute of Technology, qui n’a pas participé à la nouvelle étude. Les animaux sont si petits, note-t-il, que les minuscules griffes de leurs pattes font environ un dixième de la largeur d’un cheveu humain.

10. Only amber can preserve tardigrades in such minute detail.

a. Seul l’ambre permet de conserver les tardigrades aussi intacts.
b. Seul l’ambre peut préserver les tardigrades avec autant de détails.
c. Seul l’ambre peut préserver les tardigrades avec un tel niveau de détail minutieux.
d. Seul l’ambre peut préserver les tardigrades avec un tel détail.

11. The rarity of tardigrade fossils, however, is not just attributable to their tiny size.

a. La rareté des fossiles de tardigrades n'est toutefois pas uniquement due à leur taille minuscule
b. La rareté des fossiles de tardigrades n’est cependant pas uniquement attribuable à leur petite taille.
c. La rareté des fossiles de tardigrades n'est toutefois pas seulement due à leur petite taille
d. Toutefois, si les fossiles de tardigrades sont rares, ce n’est pas seulement à cause de leur taille.

12. Just a few paleontologists study fossil tardigrades, Mapalo says, noting that some colleagues react with surprise that any fossil tardigrades are known at all.

a. « Seuls quelques paléontologues étudient les fossiles de tardigrades », indique Marc Mapalo, en précisant que « certains de ses collègues sont même surpris que des fossiles de tardigrades puissent exister ».
b. Seuls quelques paléontologues étudient les fossiles de tardigrades, dit Mapalo, notant que certains de ses collègues sont surpris d'apprendre que des fossiles de tardigrades existent même.
c. Seuls quelques paléontologues étudient les tardigrades fossiles, explique Mapalo, notant que certains collègues réagissent avec surprise à l’idée que des fossiles de tardigrades soient connus.
d. Seuls quelques paléontologues étudient les tardigrades fossiles, explique M. Mapalo, qui note que certains de ses collègues s'étonnent que l'on connaisse des tardigrades fossiles.

13. Modern imaging techniques can help experts to squeeze new information out of previously collected amber samples.

a. Les techniques modernes d’imagerie peuvent aider les experts à extraire de nouvelles informations des échantillons d’ambre prélevés précédemment.
b. Les techniques d'imagerie modernes peuvent aider les experts à extraire de nouvelles informations d'échantillons d'ambre collectés antérieurement.
c. Les techniques d’imagerie modernes peuvent aider les experts à recueillir de nouvelles informations à partir des échantillons d’ambre à leur disposition
d. Grâce aux techniques d’imagerie moderne, les experts ont pu obtenir de nouvelles informations à partir des échantillons d’ambre collectés par le passé.

14. Mapalo and his coauthors turned to a technique called confocal fluorescence microscopy to create high-resolution images of the tiny creatures. The experts found that the two fossil tardigrade species in the amber sample aren’t alive today, but both belong to tardigrade families that are still around. By comparing the Canadian fossils and two others found in New Jersey to molecular data from living species, Mapalo and his colleagues were able to estimate when tardigrades evolved and when they gained one of their most remarkable abilities.

a. Mapalo et ses co-auteurs ont utilisé une technique appelée microscopie confocale à fluorescence pour créer des images haute résolution des minuscules créatures. Les experts ont découvert que les deux espèces fossiles de tardigrades dans l'échantillon d'ambre ne sont plus vivantes aujourd'hui, mais appartiennent toutes deux à des familles de tardigrades encore existantes. En comparant les fossiles canadiens et deux autres trouvés dans le New Jersey à des données moléculaires d’espèces vivantes, Mapalo et ses collègues ont pu estimer quand les tardigrades ont évolué et quand ils ont acquis l'une de leurs capacités les plus remarquables.
b. Marc Mapalo et ses co-auteurs ont eu recours à la technique de la microscopie confocale à fluorescence, qui leur a permis d’obtenir des images haute résolution des petites créatures. Ils ont pu déterminer que les deux espèces de tardigrades fossilisés dans l’ambre n’étaient plus vivantes, mais qu’elles

appartenaient à des familles encore présentes sur Terre. En comparant les données moléculaires d'espèces vivantes aux fossiles canadiens et à deux autres provenant du New Jersey, l'équipe de chercheurs est parvenue à estimer la date à laquelle les tardigrades ont évolué et ont acquis une de leurs incroyables particularités.
c. Mapalo et ses coauteurs se sont tournés vers une technique appelée microscopie confocale à fluorescence pour créer des images haute résolution des minuscules créatures. Les experts ont constaté que les deux espèces fossiles de tardigrades dans l'échantillon d'ambre ne sont pas encore vivantes, mais qu'elles appartiennent toutes deux à des familles de tardigrades qui existent toujours. En comparant les fossiles canadiens et deux autres trouvés dans le New Jersey à des données moléculaires d'espèces vivantes, Mapalo et ses collègues ont pu estimer quand les tardigrades ont évolué et quand ils ont acquis l'une de leurs capacités les plus remarquables.
d. Mapalo et ses coauteurs se sont tournés vers une technique appelée microscopie à fluorescence confocale pour créer des images haute résolution des minuscules créatures. Les experts ont découvert que les deux espèces de tardigrades fossiles présentes dans l'échantillon d'ambre ne sont pas vivantes aujourd'hui, mais qu'elles appartiennent toutes deux à des familles de tardigrades qui existent encore. En comparant les fossiles canadiens et deux autres découverts dans le New Jersey aux données moléculaires d'espèces vivantes, Mapalo et ses collègues sont parvenus à estimer à quelle période les tardigrades ont évolué et quand ils ont acquis l'une de leurs capacités les plus remarquables.

15. Many tardigrades are capable of cryptobiosis, a temporary and almost complete slowdown of their bodies' processes. In this state of suspended animation, the creatures shed their water and curl into balls. Along with carrying a protein that protects their DNA from damage, being able to shut down and wait for better conditions helped tardigrades to survive in extreme environments, even the vacuum of space, and could help them withstand a future apocalypse.

a. De nombreux tardigrades sont capables de cryptobiose, une extinction temporaire quasiment complète des processus corporels. Dans cet état de vie interrompue, ces créatures se vident de l'eau qu'elles contiennent et se roulent sur elles-mêmes. Outre la protéine dont les tardigrades disposent et qui protège leur ADN de toute dégradation, l'aptitude à stopper leurs processus corporels dans l'attente de conditions plus favorables leur a permis de survivre dans des environnements extrêmes, y compris dans le vide de l'espace, et pourrait même les aider à résister à une éventuelle apocalypse.
b. De nombreux tardigrades sont capables de cryptobiose, un ralentissement temporaire et presque complet des processus de leur corps. Dans cet état d'animation suspendue, les créatures perdent leur eau et se recroquevillent en boule. En plus de porter une protéine qui protège leur ADN des dommages, la capacité de s'arrêter et d'attendre de meilleures conditions a aidé les tardigrades à survivre dans des environnements extrêmes, même dans le vide spatial, et pourrait les aider à résister à une future apocalypse.
c. De nombreux tardigrades sont capables d'entrer en cryptobiose, un ralentissement temporaire et presque total de leur métabolisme. Dans cet état d'arrêt temporaire des fonctions vitales, les créatures expulsent l'eau contenue dans leur corps et se recroquevillent en boule. En plus de porter une protéine qui protège leur ADN des dommages, les tardigrades sont capables de se mettre en veille en attente de jours meilleurs, ce qui leur a permis de survivre à des environnements extrêmes et même au vide spatial, et pourrait les aider à résister à un futur apocalypse.
d. De nombreux tardigrades sont capables de cryptobiose, un ralentissement temporaire et presque total des processus de leur corps. Dans cet état d'animation suspendue, les créatures se débarrassent de leur eau et se mettent en boule. En plus d'être porteurs d'une protéine qui protège leur ADN des dommages, les tardigrades sont capables de s'arrêter et d'attendre de meilleures conditions, ce qui leur permet de survivre dans des environnements extrêmes, même dans le vide spatial, et pourrait les aider à résister à une future apocalypse.

16. Mapalo and colleagues propose that at least two major tardigrade groups evolved their cryptobiotic abilities independently, one gaining cryptobiosis between 430 and 175 million years ago and another doing so between 382 and 175 million years ago.

a. Mapalo et ses collègues suggèrent qu'au moins deux grands groupes de tardigrades ont développé leurs capacités cryptobiotiques de manière indépendante, l'un ayant acquis la cryptobiose il y a entre 430 et 175 millions d'années et l'autre entre 382 et 175 millions d'années.
b. D'après Mapalo et ses collègues, au moins deux groupes de tardigrades majeurs ont développé la capacité de cryptobiose de façon indépendante, l'un entre 430 et 175 millions d'années et l'autre entre 382 et 175 millions d'années avant notre ère.
c. D'après Marc Mapalo et ses collègues, au moins deux grands groupes de tardigrades ont développé des capacités de cryptobiose chacun de leur côté : l'un il y a 175 à 430 millions d'années, et l'autre il y a 175 à 382 millions d'années.
d. Selon Marc Mapalo et ses collègues, au moins deux grands groupes de tardigrades ont développé des aptitudes cryptobiotiques de façon indépendante. Le premier serait devenu capable de cryptobiose il y a entre 430 et 175 millions d'années et l'autre entre 382 et 175 millions d'années.

17. More fossils could help refine the exact timing, but the researchers note that this span of prehistoric time is significant because it includes

several mass extinctions. Tardigrades that were able to go into a form of stasis until conditions recovered would have been better able to survive the oxygen drops, climate shifts, and other pressures associated with these global disasters.
a. L'analyse d'autres fossiles pourrait permettre d'affiner la chronologie exacte, mais les chercheurs constatent que cette période préhistorique est importante, car elle comprend plusieurs extinctions de masse. Les tardigrades qui ont pu entrer dans une forme de stase en attendant que les conditions s'améliorent auraient été mieux à même de survivre au manque d'oxygène, aux changements climatiques et aux autres pressions associées à ces catastrophes planétaires.
b. D'autres fossiles pourraient affiner cette chronologie, mais les chercheurs notent que cette période préhistorique est significative car elle inclut plusieurs extinctions massives. Les tardigrades capables de se mettre en stase jusqu'à ce que les conditions s'améliorent auraient eu une meilleure chance de survivre aux baisses d'oxygène, aux changements climatiques et à d'autres pressions liées à ces catastrophes mondiales.
c. Davantage de fossiles pourraient aider à préciser le moment exact, mais les chercheurs notent que cette période préhistorique est importante parce qu'elle comprend plusieurs extinctions massives. Les tardigrades qui ont été capables de passer sous une forme de stase jusqu'à ce que les conditions récupérées auraient mieux pu survivre aux chutes d'oxygène, aux changements climatiques et aux autres pressions associées à ces catastrophes mondiales.
d. D'autres fossiles pourraient aider à préciser la chronologie exacte, mais les chercheurs notent que cette période préhistorique est importante car elle comprend plusieurs extinctions massives. Les tardigrades qui ont pu entrer dans une forme de stase jusqu'à ce que les conditions se rétablissent auraient été mieux à même de survivre aux baisses d'oxygène, aux changements climatiques et aux autres pressions associées à ces catastrophes mondiales.

18. "Knowing when cryptobiosis evolved in tardigrades can help us contextualize how and why they gained this mechanism," Mapalo says. Tardigrades likely evolved in the seas before spreading onto land, he notes.
a. « Savoir quand la cryptobiose a évolué chez les tardigrades peut nous aider à comprendre comment et pourquoi ils ont acquis ce mécanisme », explique M. Mapalo. Les tardigrades ont probablement évolué dans les mers avant de se répandre sur la terre ferme, note-t-il.
b. « Savoir quand les tardigrades ont acquis la capacité d'entrer en cryptobiose peut nous aider à contextualiser comment et pourquoi ils ont développé ce mécanisme, précise Mapalo. Les tardigrades ont probablement évolué en milieu marin avant de s'aventurer sur la terre ferme. »
c. « Savoir quand la cryptobiose a évolué dans les tardigrades peut nous aider à contextualiser comment et pourquoi ils ont acquis ce mécanisme », explique Mapalo. Les tardigrades ont probablement évolué dans la mer avant de se propager sur terre, note-t-il.
d. « En comprenant quand les tardigrades ont développé la cryptobiose, nous pouvons formuler des hypothèses sur la manière et la raison de l'apparition de ce mécanisme », explique Marc Mapalo. Il est possible que les tardigrades aient évolué dans les océans avant de se répandre sur la terre ferme.

19. Cryptobiotic abilities would have helped tardigrades survive changes in salt levels when they moved from the marine realm to habitats full of mosses and lichens that relied on freshwater.
a. La cryptobiose aurait alors permis aux tardigrades de survivre aux changements de taux de salinité lorsqu'ils sont passés du monde marin aux milieux riches en mousses et en lichens qui nécessitent de l'eau douce.
b. Grâce à la cryptobiose, ils auraient pu quitter le milieu marin et s'adapter à des habitats non salés où l'eau douce faisait pousser mousses et lichens.
c. Les capacités de cryptobiose auraient pu aider les tardigrades à survivre aux changements de salinité lors du passage de l'environnement marin à un habitat composé de mousses et de lichens qui eux, dépendent de l'eau douce.
d. Les capacités cryptobiotiques des tardigrades les auraient aidés à survivre aux changements des niveaux de sel lorsqu'ils ont quitté le milieu marin pour des habitats pleins de mousses et de lichens qui se développent dans l'eau douce.

20. How exactly cryptobiosis played into the survival and evolutionary history of water bears will need more research to confirm.
a. Il faudra davantage de recherches pour confirmer le rôle exact de la cryptobiose dans la survie et l'histoire évolutive des ours d'eau.
b. Comment exactement la cryptobiose a influencé la survie et l'histoire évolutive des oursons d'eau nécessitera davantage de recherches pour être confirmée.
c. Des recherches supplémentaires devront être menées pour déterminer le rôle exact de la cryptobiose dans la survie et l'évolution des oursons d'eau.
d. D'autres recherches seront nécessaires pour confirmer le rôle exact de la cryptobiose dans l'histoire de la survie et de l'évolution des oursons d'eau.

The Challenge of Translating Culture-Specific Items: Evaluating MT and LLMs Compared to Human Translators

Bojana Budimir
University of Belgrade, Falulty of Philology / Studentski trg 3
boj.budimir@gmail.com

Abstract

We evaluate state-of-the-art Large Language Models (LLM's) ChatGPT-4o, Gemini 1.5 Flash, and Google Translate, by focusing on the translation of culture-specific items (CSIs) between an underrepresented language pair: the Flemish variant of Dutch and Serbian. Using a corpus derived from three Flemish novels we analyze CSIs in three cultural domains: Material Culture, Proper Names, and Social Culture. Translation strategies are examined on a spectrum that goes from conservation to substitution. Quantitative analysis explores strategy distribution, while qualitative analysis investigates errors, linguistic accuracy, and cultural adaptation. Despite advancements, models struggle to balance cultural nuances with understandability for the target readers. Gemini aligns most closely with human translation strategies, while Google Translate shows significant limitations. These findings underscore the challenges of translating CSIs—particularly Proper Names—in low-resource languages and offer insights for improving machine translation models.

1 Introduction

Recent advancements in machine translation (MT) have significantly enhanced its quality and broadened its applicability, even in the domain of literary translation, an area often considered resistant to automation due to its reliance on nuance, creativity, and cultural context. Existing studies have reported varying levels of success for MT tools, with accuracy rates ranging from 44% (Fonteyne et al., 2020) to 20% (Webster et al., 2020). Several researchers have investigated the potential of machine translators pre-trained on literary texts (Matusov (2019); Kuzman et al. (2019), showing that tailored systems can improve automatic evaluation metrics for prose translations when compared to baseline models.

Beyond improvements in output quality, recent scholarship has also investigated how MT and computer-assisted translation (CAT) tools might be adapted to support the specific demands of literary translation. Hadley (2023), for instance, argues that these technologies can serve as productivity aids rather than replacements for human creativity. He identifies a range of functionalities, such as sentence length control, rhyme pattern identification, and syllable counting, that could be incorporated into CAT tools to assist translators working with poetry or stylistically marked texts. Similarly, Kolb and Miller (2022) provide empirical evidence that the tool PunCAT, designed to support the translation of puns, can stimulate and broaden the translator's pool of creative solutions, thus enhancing problem-solving in areas of high linguistic density and ambiguity.

In parallel, the impact of MT on translator creativity and reader experience has also been part of several studies. Their findings revealed that while human translations exhibit a higher degree of creativity, there is no statistically significant difference in the overall reading experience between human and post-edited machine translations (Guerberof-Arenas and Toral, 2020; Guerberof-Arenas and Toral, 2022). Furthermore, large language models (LLMs) have introduced tools capable of tackling complex tasks, such as creative writing (Gomez-Rodriguez and Williams, 2023) and poetry (Porter and Machery, 2024), expanding their potential applications.

Proceedings of Machine Translation Summit XX Volume 1, pages 455–467, June 23-27, 2025

The growing capabilities of MT tools and LLMs have led to their widespread use in various fields, including literary translation. According to a recent study conducted by the European Council of Literary Translators' Associations (CEATL, 2024), more than half (54%) of literary translators from 34 member countries occasionally use MT tools in their work, primarily for translating short passages or sentences (62%). Notably, some publishers have begun offering literary translators assignments to revise machine-translated texts, with approximately 7% of translators in Serbia reporting such requests (CEATL, 2024). This trend is further supported by publishers' emerging plans to release books translated entirely by artificial intelligence (Creamer, 2024). This trend emphasizes the growing importance of evaluating MT tools from a practical, user-oriented perspective.

Despite these advancements, one of the greatest challenges in both human and machine translation remains the accurate handling of culture-specific items (CSIs). These elements are particularly complex due to their dual role: they function within the narrative structure while carrying connotations and references to concepts often absent in the target culture. This duality makes CSIs a critical focus for evaluating MT systems. Understanding how MT tools and LLMs manage culturally bound elements provides valuable insights into their performance, particularly in literary translation, where maintaining the integrity of CSIs is crucial.

This need becomes especially apparent in light of findings by Daems (2022), who who studied the use and perceived usefulness of translation technologies by Dutch literary translators. Her research shows that many literary translators consider MT and CAT tools largely inadequate for capturing essential literary features such as style, humor, irony, and metaphor, as well as broader aspects such as context and cultural background. These perceptions underscore the persistent gap between current technological capabilities and the nuanced demands of literary translation. Therefore, examining the treatment of CSIs by MT and LLM systems not only provides a means to evaluate current performance but also reveals areas in need of targeted development, contributing to the creation of more culturally aware and context-sensitive translation technologies.

This study also addresses the challenges posed by low-resource languages, such as Serbian, which lack sufficient training data for MT systems. Serbian ranks among the least technologically developed European languages, alongside Maltese, Irish, Luxembourgish, and Bosnian, as highlighted by the ELE (European Language Equality) project (Srebnik, 2023). Furthermore, in the field of machine translation research, studies on low-resource languages often focus on their pairing with dominant global languages, such as English. By examining an underrepresented language pair, this research contributes to a deeper understanding of MT performance in less-studied linguistic contexts and offers insight into existing gaps that can inform future improvements in AI tool development.

This article investigates how contemporary MT tools, including ChatGPT-4o, Gemini 1.5 Flash, and Google Translate, handle CSIs, and how their use of translation strategies compares to those of human translators. In this study, the term machine translation (MT) is used as an umbrella term encompassing both neural machine translation (NMT) systems and large language models (LLMs). These tools were chosen to facilitate a comparison between NMT and LLM-based approaches. NMT systems, such as Google Translate and DeepL, rely on large-scale parallel corpora and are expensive to develop and maintain, which limits their coverage of less-resourced language pairs. At the time of writing, Google Translate is the only major NMT service that supports translation between Dutch and Serbian, restricting access to high-quality NMT for Serbian-speaking users. In contrast, LLMs are trained on vast multilingual datasets, including monolingual and non-parallel corpora, which allows them to perform translations across a broader range of language pairs, even in low-resource scenarios. Their growing adoption by professional translators, as indicated in recent surveys such as the CEATL report (2024), further underscores their relevance to translation practice.

By analyzing strategy distribution, mistranslation rates, and error patterns across cultural categories, the study evaluates the differences between these models, identifies which tool aligns most closely

with human translation, and highlights the most common types of errors in each approach.

Related work

The term culture-specific item (CSI) was introduced by Franco Aixelá to describe elements in a text that may pose challenges for translators due to their function or connotation, particularly when the referenced phenomenon does not exist in the target culture or holds a different intertextual status in the readers' cultural framework (Aixelá, 1996). Alongside this term, translation studies have proposed a variety of other terms to describe this phenomenon, such as realia (Grit, 2010; Leppihalme, 2001), cultural words (Newmark, 1988), cultural references (Olk, 2013), and cultureme (Katan, 2009). Leppihalme (2001) defines realia as lexical elements that refer to the real world outside language, making them extralinguistic phenomena. She argues that these references can lead to what she terms a cultural bump, a situation in which the reader of the target text encounters difficulty understanding a culture-bound reference because it has no equivalent in their own cultural context (Leppihalme, 1997).

Due to their extralinguistic and culture-bound nature, CSIs require the translator to possess not only bilingual proficiency but also a high degree of bicultural competence. The successful translation of CSIs demands encyclopedic cultural knowledge as well as creativity in identifying appropriate solutions. Translators can draw upon a range of strategies and procedures to address these challenges. From a macro perspective, they may choose to preserve the original CSI, a strategy associated with foreignization, or to adapt it for the target audience through domestication (Venuti, 1995). At the micro level, various procedures exist for rendering CSIs within the text, and numerous taxonomies have been developed specifically for this purpose (Aixelá, 1996; Leppihalme, 2001; Grit, 2010; Olk, 2013). However, as Olk (2013) points out, no single taxonomy can be considered universally applicable; the selection of a particular model often depends on factors such as the objectives of the study and the language pair involved. This is also the case in the present study, where a specific taxonomy was chosen based on these methodological considerations.

The selection of an appropriate translation strategy for CSIs is influenced by numerous factors. Scholars such as Newmark (1988), Aixelá (1996), and Grit (2010) have identified patterns in the application of strategies based on the type of CSI. In addition to the inherent characteristics of CSIs, and textual features such as canonization, markedness and relevance, supratextual, textual, and intratextual parameters play a crucial role in strategy selection. Supratextual parameters include linguistic norms, reader expectations, and publisher policies, while the function of the CSI within the text is considered intertextual (Aixelá, 1996).

While much of the research on CSIs has focused on human translation, there is growing interest in how MT tools handle CSIs. Yao et al. (2024) addressed the challenges MT systems face when translating culturally specific content. They introduced a culturally aware machine translation (CAMT) parallel corpus enriched with CSI annotations and proposed a novel evaluation metric to assess translation understandability using GPT-4. This research highlights the potential of LLMs to handle complex cultural elements while revealing areas where improvements are needed. Similarly, Pudjiati et al. (2021) explored the role of post-editing in improving machine-translated CSIs from Indonesian into English. Their findings underscore the limitations of MT systems in handling figurative language and culturally nuanced terms, emphasizing the importance of human intervention in achieving semantic accuracy and cultural fidelity.

Proper names, a subset of CSIs, have also received significant attention in MT research. Hurskainen (2013) examined the challenges associated with translating proper names, highlighting the role of tagging, rule-based disambiguation, and probability measures in resolving ambiguity. This study demonstrated how linguistic and contextual rules can improve MT accuracy when handling proper names, particularly those with dual meanings or capitalization issues.

All the above-mentioned studies collectively support the present research by providing theoretical and practical insights into the complexities of CSI translation and the evolving role of MT. By focusing on low-resource

language and evaluating specific MT models, this study builds on prior work to address gaps in understanding how machine and human translators handle CSIs across cultural categories.

2 Methodology

This study is based on the research into the translation of CSIs in Flemish literature (Budimir, 2021), extending its scope to include a comparative analysis of machine translation and human translation strategies. Specifically, it investigates the translation of CSIs across three cultural categories—Material Culture (MC), Proper Names (PN), and Social Culture (SC)—as defined by Newmark (1988). Material Culture (MC) includes items such as food, drink, and towns/housing, reflecting tangible aspects of everyday life. Proper Names (PN) encompass street names, brand names, and HoReCa (hotel, restaurant, and café) names, which often require specific adaptation to the target culture. Social Culture (SC) comprises job titles, sports and games, and leisure activities, highlighting culturally embedded practices and societal roles. These categories were selected due to the clear divergence in translation strategies applied to each. As demonstrated in Budimir (2021), translators predominantly employed orthographic adaptation and literal translation when rendering PNs, thereby opting for the conservation of these elements. In contrast, description and localization were more frequently used for items related to MC and SC, where translators tended to adapt the elements to the expectations of target readers. To further support this categorization, a Chi-Square test was conducted to assess whether there were significant differences in the strategies used by the human translators within each category. The test revealed no statistically significant differences in the distribution of translation strategies between the translators ($p = 0.066$ for PN, $p = 0.256$ for MC, and $p = 0.438$ for SC). This outcome supports the assumption that it is primarily the nature of the CSI—and not the individual translator—that influences strategic choices, thereby reinforcing the relevance of these categories for analyzing patterns of translation behavior across different cultural domains.

The research adopts a mixed-method approach. Quantitative analysis examines the distribution of translation strategies employed by machine translation models and human translators, while qualitative analysis explores errors, linguistic accuracy, and cultural nuances. The methodology comprises three phases: (1) corpus formation, (2) extraction of translation equivalents, and (3) classification of translation strategies.

2.1 Corpus Formation

The corpus for this study is derived from an existing dataset of six Flemish novels. For this research, excerpts were selected from three culturally rich novels from the original corpus: *Het verdriet van België* (*The Sorrow of Belgium*) by Hugo Claus, translated into Serbian by Ivana Šćepanović and published in 2000; *De komst van Joachim Stiller* (*The Coming of Joachim Stiller*) by Hubert Lampo; and *De helaasheid der dingen* (*The Misfortunates*) by Dimitri Verhulst. The latter two were translated by Jelica Novaković-Lopušina in 1992 and 2015, respectively. These two translators are among the most productive and prominent figures working in the field of literary translation from Dutch to Serbian.

The corpus formation process began with a predefined list of 197 CSIs, from the previous research serving as the primary units of analysis. Instances of CSIs were identified within a parallel corpus organized in an Excel sheet. Sentences containing CSIs, along with preceding and following sentences, were extracted to provide contextual information. This process resulted in a corpus containing 246 sentences, 7,087 words and 43,421 characters.

Given the character limitations of machine translation models—Google Translate (5,000 characters) and ChatGPT (4,096 characters), the corpus was divided into chunks of approximately 600 words. Consistent chunks were used across all models (ChatGPT, Gemini, and Google Translate) to ensure comparability. ChatGPT produced two translation variants: ChatGPT (1), without the search option, and ChatGPT (2), with the search option. Translations were generated on November 12th 2024 using a standardized zero-shot prompt: "Translate this text from Dutch to Serbian." The use of a simple, zero-shot prompt was intentional, as the goal of the study was to evaluate the baseline performance of two LLMs and an NMT system when translating CSIs.

Strategies	Description
Repetition (R)	The CSI is kept in its original form. For Serbian, this may include adding inflectional suffixes to align with the target language's grammatical rules. For example: In de Volkskring - U *De Volkskring-u*.
Orthographic Adaptation (OA)	The CSI is adapted to reflect its pronunciation in the target language, following Serbian orthographic conventions. For example: Scheldewindeke - *Sheldevindeke.*
Combination of strategies (COM)	The CSI is either retained in its original or adapted form and supplemented with additional information, such as a classifier, an explanation integrated into the text, or a footnote. For example: De Leie - reka Leja [the river Leie].
Literal Translation (LT)	A word-for-word translation of a concept that may be unfamiliar in the target culture, preserving the source language's structure as closely as possible. For example: Het Hoekske - *Ćošak* [The Corner].
Description (D)	The CSI is replaced by a descriptive phrase or explanation to convey its meaning or function in the target language. For example: Glas-in-lood - *Okna u raznobojnom staklu* [pane with colorful glass].
Generalization (G)	The CSI is replaced with a neutral or broader reference that lacks cultural specificity. For example: Boterkoek - *Pecivo* [Pastry].
Localization (L)	The CSI is replaced with a reference specific to the target culture, making it more familiar to the target audience. For example: Hutsepot - *Čušpajz*.
Mistranslation (Mis)	Errors in translation, including incorrect orthographic adaptation, grammatical or semantic inaccuracies, or the use of non-existent words. For example: Vogelpik - *Kljucanje ptica* [Birds pecking].

Table 1: Adapted Taxonomy of Translation Strategies for Rendering CSIs. (Budimir 2021)

Varying the prompts would have introduced an additional variable, potentially influencing the outcome and making cross-model comparisons (NMT and LLMs) less meaningful.

It is important to note that dividing the corpus into chunks may disrupt the narrative flow, potentially limiting the models' ability to fully comprehend and translate context-dependent CSIs. Despite this limitation, the approach ensures consistency and accommodates the technical constraints of the models.

2.2 Translation Equivalents and Strategies

Translation equivalents were extracted in an Excel sheet, paired with the original CSI and the human translation from the previous study. On average, 206 equivalents per model were identified, as multiple translations of the same CSI were recorded.

The translations were categorized by the author, an experienced translator and researcher in the field of translation studies, using a taxonomy adapted from Budimir (2021), which includes the following strategies: Repetition (R), Orthographic Adaptation (OA), Combination of Strategies (COM), Literal Translation (LT), Description (D), Generalization (G), Localization (L), and Mistranslation (Mis). Table 1 provides a detailed description of these strategies. The taxonomy facilitates a granular analysis, capturing the diversity of approaches employed by the models and enabling meaningful comparisons with human translations.

It is necessary to point out that the classification process is inherently subjective. As the categorization was performed by a single annotator, the results may reflect one individual's interpretive biases. While the taxonomy provides clear guidelines, subjective judgment is often required to determine the most appropriate category for each translation equivalent. This limitation is particularly relevant for studies involving smaller language pairs, where finding annotators can be challenging. Future studies could mitigate this limitation by employing multiple annotators and calculating inter-annotator agreement to enhance reliability and validity.

3 Results

3.1 General Overview

A visual illustration of the distribution of translation strategies employed by the MT tools (Google Translate, Gemini, ChatGPT (1), and

ChatGPT (2)), as well as by the human translator is presented in Figure 1. A Chi-Square test was conducted to assess differences in translation strategies among the models, revealing significant variation ($\chi^2 = 100.83$, $p < 0.05$, dof = 24). These results indicate that each model exhibits distinct approaches to handling CSIs.

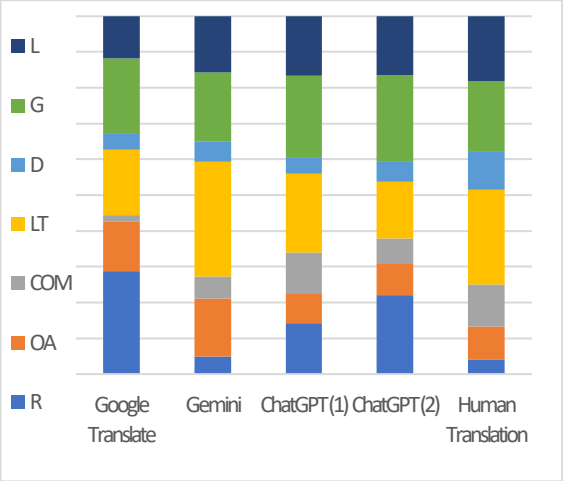


Figure 1: Distribution of Strategies across Models and Human Translators.

Mistranslation (Mis) rates (Table 2) highlight differences in model accuracy. Google Translate exhibits the highest error rate, with 56 instances (26.8%) of mistranslation, reflecting substantial challenges in handling CSIs. This result aligns with the findings of Yao et al. (2024), which demonstrate the superior ability of LLMs over NMT systems in managing CSIs. The relatively high error rate of ChatGPT (2), however, can be attributed to a specific translation issue: the omission of the case suffix in retained CSIs. This issue will be further discussed in the section 3.2.

Excluding mistranslations, the distribution of correct strategies provides insights into the strengths and weaknesses of each model. Google Translate relies heavily on Repetition (R) (21.1%), indicating a tendency to preserve CSIs in their original form without adaptation. This approach often contradicts Serbian norms, where orthographic adaptation is preferred. In contrast, ChatGPT (both versions) exhibits the strongest reliance on Generalization (G) and Localization (L), reflecting an effort to adapt cultural references for the target audience. ChatGPT (1)'s frequent use of the Combination of Strategies (COM) underscores its capacity to enhance contextual clarity, while Google Translate and Gemini rely more on straightforward strategies, offering limited additional explanation.

Among machine translation models, Gemini demonstrates the most balanced distribution of strategies. It effectively integrates Literal Translation (LT), Orthographic Adaptation (OA), Generalization (G) and Localization (L), suggesting a more adaptable approach. This balance mirrors the diversity observed in human translation more closely than in either Google Translate or ChatGPT, which exhibit a narrower range of strategies. Statistical metrics support this conclusion, as demonstrated by measuring Euclidean distance—a method commonly used to evaluate similarity between categorical data—between each model and human translation as the baseline. Gemini exhibits the smallest Euclidean distance from human translation (0.122), followed by ChatGPT (1) (0.133). ChatGPT (2) (0.227) and Google Translate (0.297) display greater divergence.

The analysis of the distribution of strategies across cultural categories (Figure 2) offers additional insights. ChatGPT (2) and Gemini produce the highest number of errors in the Proper Names (PN) category, indicating significant challenges in adapting names to Serbian linguistic norms. In contrast, ChatGPT (1) and Google Translate show the most errors in the Social Culture (SC) category, suggesting difficulties in handling references to job titles, leisure activities, and institutions. These results highlight how different models struggle with specific cultural categories, reflecting varying capabilities in adapting to cultural and linguistic nuances. Google Translate continues to demonstrate a relatively consistent struggle across all categories, underlining its limited cultural sensitivity.

For Material Culture (MC) references, all machine translation models frequently rely on Generalization and Literal Translation. Human Translation, by contrast, employs Generalization (28%) alongside a stronger preference for Description (24%) and Literal Translation (18.7%).

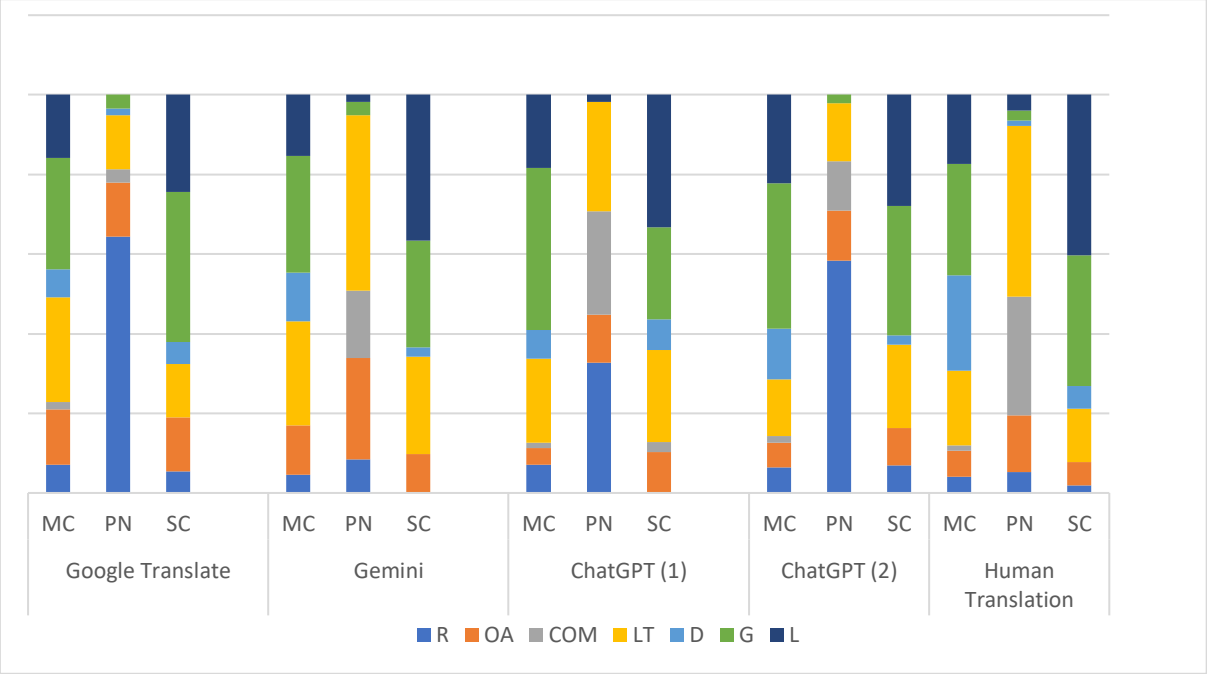


Figure 2: Distribution of Strategies across Cultural Categories and Models.

For Proper Names (PN), Repetition is predominant in Google Translate, while Gemini and ChatGPT favor Literal Translation. Human Translation demonstrates a preference for Literal Translation (42.9%) and the Combination of Strategies (29.9%), reflecting its emphasis on adapting to the target culture and providing contextual information. In the case of Social Culture, Generalization and Localization are dominant strategies for both machine models and Human Translation. However, machine models employ Localization less frequently (e.g. 16.5% in ChatGPT (2)) than Human Translation (40.4%).

Human Translation consistently prioritizes Localization and Generalization for Social and Material Culture, while favoring Literal Translation for Proper Names, thus demonstrating a clear preference for culturally adaptive strategies. In contrast, machine models show less consistency and rely more heavily on Generalization and Literal Translation, particularly in more challenging categories.

3.2 Error Analysis

Semantic errors were the most prevalent errors across all models (Figure 3), reflecting the significant challenges these systems face with the polysemy of multi-word expressions and exocentric compound words, which CSIs often comprise. Insights from studies on polysemous words emphasize the importance of contextual dependency in resolving such errors. Machine translation often fails to disambiguate polysemous terms and uses primary meanings without considering context (Ohuoba et al., 2024). For instance, Google Translate rendered *jarige kaas* as *rođendanski sir* [birthday cheese], incorrectly interpreting *jarig* as "birthday" rather than its actual meaning in this context, "aged"—the correct translation being "aged cheese". Similarly, *zure spekken* was mistranslated as *kisele slanine*

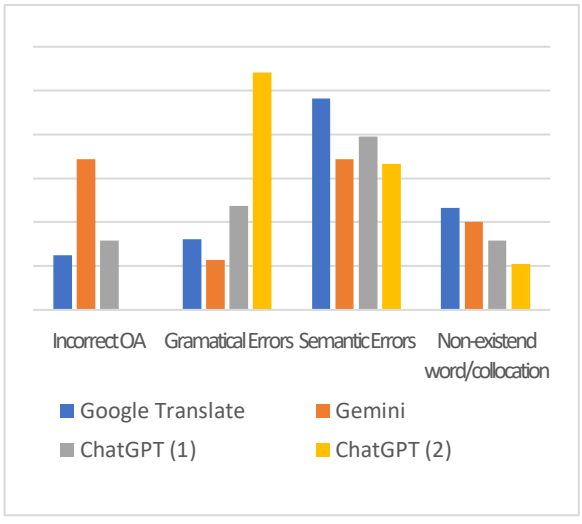


Figure 3: Distribution of Error Occurrences (%) across Models.

[sour bacon], where *spekken* (a type of soft candy) was wrongly interpreted as "bacon".

After Google Translate (23.3%), Gemini exhibited the highest frequency of invented words and collocations (20%), with examples including *takmičenje u kašetanju* for *kaatwedstrijd* [a sort of ball sport] and *kafić Korabljavanje* for *cafe De Scheepvaart* [cafe Shipping]. These outputs suggest a tendency to generate nonsensical and non-existent terms in Serbian, reflecting lexical gaps in the model's training data and its resorting to hallucination. While ChatGPT produced fewer non-existent terms, similar issues were observed, indicating room for improvement in vocabulary alignment with Serbian norms.

Grammatical errors were particularly pronounced in ChatGPT (2), with a notable 54.2% rate, largely due to challenges with Serbian case endings, agreement, and word order (16 out of 26 errors in this category). For example, *Grote Markt* [Main Square] was translated without the proper locative case ending (*na Grote Markt* instead of *Markt-u*), and brand names were often repeated without morphological adaptation, as in *flaša Bols likera* instead of *flaša Bolsovog likera* [bottle of Bols liqueur]. One prominent error in Gemini involved the incorrect use of the suffix *-ski* instead of *-ov* when forming adjectives from people's names. For instance, *Snellaertstraat* was incorrectly adapted as *Snerlatska ulica* instead of the correct *Snelartova ulica*.These errors disrupt the syntactic and morphological coherence of the output, diminishing its overall accuracy.

Incorrect orthographic adaptation (OA) was another common issue, especially in Gemini (34.3%) and ChatGPT (1) (15.8%). For example, place and street names were often inconsistently adapted, violating Serbian orthographic norms. For example, the diphthong *ui* is adapted as *u* in *Oostduinkerke* and as *iu* in *Korte Gasthuisstraat.* The correct form should be *aj*. In contrast, ChatGPT (2) avoided such errors entirely, due to the predominant use of Repetition.

As highlighted in the Prolex study on French, Serbian, and Bulgarian (Maurel et al., 2007), rich inflectional systems require proper names to be adapted across cases (e.g., nominative, genitive, dative), which adds complexity to translation tasks. Serbian proper names exhibit multiple inflectional forms, underscoring the need for MT systems to incorporate morphological rules effectively.

The influence of English was most apparent in Google Translate's outputs. For example, the café name *Het Hoekske* was translated into Serbian with the English definite article "the," resulting in *The Hoekske*. Similarly, the term *bloedworst*, a type of blood sausage commonly found in Flemish cuisine, was incorrectly rendered as *crni puding* [black pudding], borrowing the literal English term, which does not align with typical Serbian culinary terminology. Another case is the translation of *schorseneer* (a root vegetable known as salsify) where the English term salsify was transferred directly into Serbian. These examples illustrate the system's reliance on English as an intermediary language, which can distort meaning and reduce the cultural and linguistic accuracy of the target text. These findings are consistent with the challenges described by Ohuoba et al. (2024), where English's dominance as a high-resource language often skews translations for low-resourced languages, by introducing cultural mismatches and semantic inaccuracies.

Some translations included lexical forms from closely related languages such as Croatian or Slovenian, such as *vrtuljak* for *paardenmolen* (a Flemish term for "carousel") and *pivovarna* for brouwerij (brewery). While these forms may be intelligible to Serbian speakers due to the linguistic similarities among South Slavic languages, they are less common or non-standard in Serbian. This highlights potential inconsistencies in the models' adaptation to regional language norms and raises questions about the influence of neighboring languages on machine-generated output.

4 Discussion

The present analysis of translation strategies and error patterns highlights the key challenges of machine translation of CSIs and especially into morphologically complex languages like Serbian. When compared to human translation, machine translation exhibits significant problems in handling CSIs.

The study of human translation strategies reveals clear and consistent patterns when dealing with proper names, including street names, brand names, and HoReCa terminology (Budimir, 2021). Orthographic adaptation is commonly applied when the name includes people's names, particularly historical figures, well-known fictional or real people, or toponyms. In contrast, literal translation is used when the name consists of nouns and adjectives. This systematic approach ensures that cultural and semantic nuances are preserved in the target text while maintaining the readability and cultural familiarity for the audience. Machine translation, however, fails to follow such patterns, often producing random and inconsistent results.

Another notable issue is the predominant use of repetition without adding contextual information. While repetition can sometimes suffice when the meaning of the CSI can be inferred from context, this is often not the case. The preservation of the communicative function is a crucial aspect of translating CSIs (Ivir, 2003). Simply repeating a term without adaptation or explanation can fail to convey its intended cultural significance, leaving the target audience disconnected from the original message. For instance, retaining pastries such as *mastellen* and *pistoletten* in their original forms does not evoke any cultural or semantic associations apart from the act of eating. This approach neglects the cultural connotations and traditional significance attached to these items in the source culture. Such CSIs require additional strategies, such as adaptation or explanatory supplementation, to ensure that their cultural and communicative essence is effectively conveyed (Ivir, 2003). Without added context, the meaning and significance of these items are lost to the Serbian audience, reducing the overall effectiveness of the translation (Hlebec, 2009).

5 Conclusion

This study has highlighted several key findings regarding the performance of machine translation (MT) systems in translating culture-specific items (CSIs) between Flemish Dutch and Serbian. First, while models such as Gemini and ChatGPT demonstrate a promising use of generalization and localization strategies for material and social culture CSIs, they often fail to apply nuanced approaches required for complex or less common CSIs. Proper names, in particular, pose significant challenges due to the rich inflectional demands of Serbian and the need for orthographic adaptation.

From a strategy perspective, the analysis reveals that Gemini exhibits the most balanced distribution of approaches, incorporating literal translation, orthographic adaptation, generalization, and localization more effectively than other models. Nevertheless, even Gemini struggles with systematic cultural adaptation and fails to match the nuanced strategies consistently employed in human translation. ChatGPT's use of the combination strategies shows potential for improving contextual clarity, yet its tendency to omit morphological adaptations in Serbian remains a limitation. Meanwhile, Google Translate, while heavily reliant on repetition, exhibits the highest error rates and demonstrates limited cultural sensitivity in handling CSIs.

These findings underscore the irreplaceable role of human translators in effectively handling CSIs, particularly in literary and culturally rich texts. Human translators not only bring cultural and contextual understanding to the task, but also excel at preserving the communicative function of CSIs, a dimension often overlooked by MT systems. For example, while MT models tend to rely on repetition or overgeneralization, human translators adapt CSIs dynamically, ensuring that their cultural essence and intended meanings resonate with the target audience.

Furthermore, the implications of this study extend to translator training and workflow design. As MT systems become more prevalent, human translators are increasingly assuming roles as post-editors. This shift emphasizes the importance of equipping translators with the skills needed to identify and address the shortcomings of MT outputs, such as the failure to capture cultural nuances or apply morphological adaptations. By integrating human expertise with MT capabilities, translation workflows can achieve greater efficiency while preserving linguistic and cultural fidelity.

Recent studies have increasingly emphasized the potential of CAT and MT tools to enhance translator efficiency, particularly when dealing with complex or culture-bound elements that require extensive background research and

strategic decision-making. Hadley (2023) highlights how such tools can alleviate cognitive load by supporting specific aspects of literary translation, such as managing sentence length, rhythm, or poetic form. Similarly, Kolb and Miller (2022) demonstrate that specialized tools like PunCAT can aid in resolving linguistically dense challenges, such as puns, by expanding the translator's pool of potential solutions. These developments suggest promising avenues for future tool design.

In the context of CSI translation, where human translators often invest significant time in interpreting meaning and selecting appropriate strategies, MT systems could be further adapted to present a range of contextually informed suggestions. Experimenting with prompt engineering, designed to generate multiple culturally and linguistically relevant options for each CSI, may prove especially beneficial in supporting informed and efficient human decision-making. In this regard, hybrid human-machine approaches and the development of culturally aware translation tools are crucial. Yao et al. (2024) provide a compelling framework for advancing MT by integrating cultural databases and CSI annotations, as well as introducing innovative metrics to evaluate cultural and contextual fidelity. Building on such approaches could significantly enhance MT performance, particularly for texts with rich cultural content.

One concrete avenue for such improvement involves addressing the persistent errors in orthographic adaptation, particularly when translating into morphologically rich languages like Serbian. These errors could be mitigated through the integration of language-specific orthographic rules, culturally adapted name databases, and targeted post-editing support within LLM systems. Such refinements, combined with the insight and flexibility of human translators, would allow for more accurate and culturally resonant translations of proper names and other CSIs.

Limitations

It should be noted that the present study is limited by its relatively small dataset of 197 analyzed CSIs, which restricts the generalizability of its conclusions. Additionally, the reliance on a single annotator introduces potential subjectivity in strategy classifications. To address these limitations, future research should analyze larger datasets, employ multiple annotators for improved reliability, and explore different datasets, including other low-resource languages, to test the consistency of observed patterns. Investigating the impact of varying prompts for large language models (LLMs) and experimenting with hybrid approaches that combine machine translation and human post-editing could further enhance the understanding and handling of culturally nuanced content. Such advancements would contribute to more robust cultural adaptation and contextual modeling in MT systems, aligning them more closely with human translation standards.

Ethics Statement

The corpus used in this research has been utilized exclusively for academic and research purposes, in compliance with copyright and ethical guidelines. All texts within the corpus have been accessed and processed solely to analyze translation strategies and linguistic phenomena as part of this study.

The corpus is securely stored on a private computer and is not accessible on any online platform or public repository. No part of the corpus has been shared, distributed, or made available beyond the scope of this research.

For the purposes of analysis, the texts were processed as loose, decontextualized sentences to focus on specific translation patterns and strategies. This approach ensures that the study adheres to ethical research standards while minimizing potential risks associated with handling copyrighted material in its entirety.

Researchers interested in the corpus for academic purposes may request access by contacting the author directly, subject to appropriate ethical and copyright considerations.

Sustainability Statement

The estimated energy usage for this study was negligible, as it involved text processing and analysis rather than high-resource training or large-scale inference tasks. As such, the environmental impact of the research is minimal.

References

Primary sources

Hugo Claus. 1983. *Het verdriet van België*. De Bezige Bij, Amsterdam.

Hugo Klaus. 2000. *Tuga Belgije T.1 Tuga*. Prometej, Novi Sad.

Hubert Lampo. 1960. *De komst van Joachim Stiller*. Meulenhoff, Amsterdam.

Hubert Lampo. 1992. *Dolazak Joahima Štilera*. Luta, Beograd.

Dimitri Verhulst. 2006. *De helaasheid der dingen*. Contact, Amsterdam.

Dimitri Verhulst. 2015. *Zaludnost življenja*. Clio, Beograd.

Secondary sources

Javier Franco Aixelà. 1996. Culture-Specific Items in Translation. In R. Álvarez and M. C. Vidal, editors, *Translation, Power, Subversion.* Multilingual Matters, Philadelphia, pages 52–78.

Ana Guerberof-Arenas, and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, 9(2):255–282.

Ana Guerberof-Arenas, and Antonio Toral. 2022. Creativity in translation: machine translation as a constraint for literary texts. *Translation Spaces*, 11(2): 184-212.

Бојана Будимир. 2021. Културноспецифични елементи из фламанске културе у преводу на српски језик [*Culture-Specific Items from Flemish Culture in Serbian Translation*]. D.Phil. dissertation, University of Belgrade, Faculty of Philology, Belgrade, Serbia.

CEATL. 2024. AI Survey for individual translators. https://www.ceatl.eu/wp-content/uploads/2024/04/CEATL_AI_survey_for_members.pdf.

Ella Creamer. 2024. Dutch publisher to use AI to translate 'limited number of books' into English. *Guardian*.

Joke Daems. 2022. Dutch literary translators' use and perceived usefulness of technology: The role of awareness and attitude. In *Using technologies for creative-text translation*. Routledge, pages 40-65.

Margot Fonteyne, Arda Tezcan, and Lieve Macken. 2020. Literary Machine Translation under the Magnifying Glass: Assessing the Quality of an NMT-Translated Detective Novel on Document Level. *Language Resources and Evaluation*, 3790–3798. https://biblio.ugent.be/publication/8662553/file/8662566.pdf

Diederik Grit. 2010. De vertaling van realia. In T. Naaijkens, editor, *Denken over vertalen*. Vantilt, Nijmegen, pages 189–196.

Vladimir Ivir. 2003. Translation of Culture and Culture of Translation. *SRAZ XLVII-XLVIII***:** 117–126.

James Luke Hadley. 2023. MT and CAT: Challenges, Irrelevancies, or Opportunities for Literary Translation?. In *Computer-assisted Literary translation*. Routledge, pages 91-105.

Борис Хлебец. 2009. *Опита начела превођења* [*General Principles of Translating*]. Београдска књига, Београд.

Arvi Hurskainen. 2013. Handling proper names in Machine Translation. Technical Report 12.

David Katan 2009. Translation as intercultural communication. In J. Munday, editor, *The Routledge Companion to Translation Studies.* Routledge, London/New York, pages 74–92.

Waltraud Kolb, and Tristan Miller. 2022. Human–computer interaction in pun translation. In *Using technologies for creative-text translation*. Routledge, pages 66-88.

Taja Kuzman, Špela Vintar, and Mihael Arčan. 2019. Neural Machine Translation of Literary Texts from English to Slovene. In *Proceedings of the Qualities of Literary Machine Translation*, pages 1–9, Dublin, Ireland. European Association for Machine Translation.

Ritva Leppihalme. 1997. *Culture bumps: An empirical approach to the translation of allusions*. Multilingual Matters.

Ritva Leppihalme. 2001. Translation strategies for realia. In *Mission, vision, strategies, and values: a celebration of translator training and translation studies in Kouvola*. Helsinki University Press, pages 139-148.

Evgeny Matusov. 2019. The Challenges of Using Neural Machine Translation for Literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.

D. Maurel, D. Vitas, C. Krstev, S. Koeva. 2007. Prolex: a lexical model for translation of proper names. Application to French, Serbian and Bulgarian. In A. Dziadkiewicz and I. Thomas, editors, *Bulag - Bulletin de Linguistique Appliquée et Générale, Les langues slaves et le français : approches formelles dans les études contrastives*, No. 32, pages 55–72, Presses Universitaires de Franche Comté, Besançon.

Peter Newmark. 1988. *A Textbook of Translation*. Prentice Hall, New York/London.

Adaeze Ohuoba, Serge Sharoff, and Callum Walker. 2024. Quantifying the Contribution of MWEs and Polysemy in Translation Errors for English–Igbo MT. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 537–547, Sheffield, UK. European Association for Machine Translation (EAMT).

Harald Martin Olk. 2013. Cultural references in translation: a framework for quantitative translation analysis. *Perspectives,* 21(3): 344-357

Brian Porter and Édouard Machery. 2024. AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Dental science reports*, 14(1).

Danti Pudjiati, Ninuk Lustyantie, Ifan Iskandar, and Tira Nur Fitria. 2022. Post-editing of machine translation: Creating a better translation of cultural specific terms. *Language Circle: Journal of Language and Literature*, 17(1): 61-73.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A Confederacy of Models: a Comprehensive Evaluation of LLMs on Creative Writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

Anita Srebnik. 2023. Het taaltechnologische landschap van het Nederlands in een meertalig Europa. *Internationale Neerlandistiek*, 61(3): 217–241.

Lawrence Venuti. 1995. The Translator's Invisibility: A History of Translation. Routledge, London/New York.

Rebecca Webster, Margot Fonteyne, Arda Tezcan, Lieve Macken, and Joke Daems. 2020. Gutenberg goes neural: comparing features of Dutch human translations with raw neural machine translation outputs in a corpus of English literary classics. *Informatics (Basel)*, 7(3):32.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking Machine Translation with Cultural Awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

A Additional Data

Table 2 presents a detailed overview of the distribution of translation strategies and mistranslation employed by MT models (Google Translate, Gemini, ChatGPT (1), ChatGPT (2)) and human translation.

	Google Translate	**Gemini**	**ChatGPT (1)**	**ChatGPT (2)**	**Human Translation**
Repetition (R)	44 (28.8%)	8 (4.8%)	24 (14.3%)	35 (22.2%)	8 (3.9%)
Orthographic Adaptation (OA)	21 (13.7)	27 (16.4%)	14 (16.4%)	14 (8.9%)	19 (9.3%)
Combination of Strategies (COM)	3 (2%)	10 (6.1%)	19 (6.1%)	11 (7%)	24 (11.8%)
Literal Translation (LT)	28 (18.3%)	53 (32.1%)	37 (32.1%)	25 (15.8%)	54 (26.5%)
Description (D)	7 (4.6%)	9 (5.5%)	8 (5.5%)	9 (5.7%)	22 (10.8%)
Generalization (G)	32 (20.9%)	32 (19.4%)	38 (19.4%)	38 (24.1%)	40 (19.6%)
Localization (L)	18 (11.8%)	26 (15.8%)	28 (15.8%)	26 (16.5%)	37 (18.1%)
Total	153	165	168	158	204
Mistranslation (Mis)	56 (26.8%)	35 (17.5%)	38 (18.4%)	48 (23.3)	1 (0.5%)

Table 2: Distribution of Strategies and Mistranslation across Models and Human Translators.

Investigating the Integration of LLMs into Trainee Translators' Practice and Learning: A Questionnaire-based Study on Translator-AI Interaction

Xindi Hao & Shuyin Zhang*
The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, P.R. China
zhangshuyin@cuhk.edu.cn

Abstract

In recent years, large language models (LLMs) have drawn significant attention from translators, including trainee translators, who are increasingly adopting LLMs in their translation practice and learning. Despite this growing interest, to the best of our knowledge, no LLM has yet been specifically designed for (trainee) translators. While numerous LLMs are available on the market, their potential in performing translation-related tasks is yet to be fully discovered. This highlights a pressing need for a tailored LLM translator guide, conceptualized as an aggregator or directory of multiple LLMs and designed to support trainee translators in selecting and navigating the most suitable models for different scenarios in their translation tasks. As an initial step towards the development of such a guide, this study aims to identify the scenarios in which trainee translators regularly use LLMs. It employs questionnaire-based research to examine the frequency of LLM usage by trainee translators, the average number of prompts, and their satisfaction with the performance of LLMs across the various scenarios identified. The findings give an insight into when and where trainee translators might integrate LLMs into their workflows, identify the limitations of current LLMs in assisting translators' work, and shed light on a future design for an LLM translator guide.

1 Introduction

Large language models (LLMs) function as the foundation models of Generative AI (GenAI) in performing text generation and language processing (Bhupathi, 2025). Very recently, the advent of LLMs has significantly impacted the translation industry. LLMs such as GPT-4, one of the latest in the Generative Pre-trained Transformer (GPT) series, BERT, and LLaMA have quickly become popular tools in translators' workstations, reshaping established practices. In translation industry, there are also translation-specific LLMs or LLM-integrated computer-assisted translation (CAT) tools, such as Trados Copilot and Wordscope, that are primarily designed for translation providers and professional translators. These AI-powered commercial tools provide professional translators with an all-in-one solution for their translation practice (Wordscope). Unlike traditional NMT which is purely an approach to automatic machine translation (Mohamed et al., 2021), with their "inherent ability to understand, generate, and manipulate human-like text in a contextually relevant manner" (Naveed et al., 2023), LLMs can be applied to a wide range of natural language processing (NLP) tasks, including question answering, summarization, text generation, and others. In other words, beyond their direct application to translation in the narrow sense, the high versatility of LLMs and their ability to be customized through prompt engineering can enable them to assist with various tasks across the entire translation workflow.

The potential of LLMs in the translation industry warrants further exploration. In modern translation services, a translation project can, by and large, be divided into three phases: pre-production, production, and post-production, as outlined in the two standards, ISO 17100:2015 and ISO 11669:2024. While these standards are designed to provide guidance for translation service providers from a project management perspective, covering various administrative activities, many of the outlined tasks are also performed by, or involve, individual translators, even during the pre- and

Proceedings of Machine Translation Summit XX Volume 1, pages 468–484, June 23-27, 2025

post-production stages. The key stages and tasks a translator may encounter throughout the entire process are summarized in Table 1, adapted from these standards, with tasks more closely aligned with managerial responsibilities excluded, as these typically fall under the role of a project manager.

Phase	Tasks / Stages
Pre-production	Setting up translation memories, terminological databases, style-guides
	Preparation of the content for translation technology processing
	Source language content analysis
	Collection and preparation of reference materials
Production	Translation
	Check
	Revision
	Review
	Proofreading
	Final verification and release
Post-production	Feedback collection

Table 1: Three Phases of a Translation Project

It is not difficult to envisage LLMs being incorporated into many of these tasks or stages. When examining the task of translation in the narrow sense, it involves several functions where LLMs might be helpful, serving, for instance, as a dictionary, as an machine translation system providing reference translations, or even as a subject-matter expert by offering domain-specific knowledge, not to mention the fact that they could potentially be applied to more complex pre-production tasks, such as content analysis and terminology extraction, as well as in post-production, where they might support feedback collection through the analysis of reviewer comments or client input.

To date, much attention has been directed to claims of human parity in the translation abilities of LLMs, with a particular focus on their performance as machine translation systems—both in terms of evaluation (Hendy et al., 2023) and improvement (Bawden & Yvon, 2023; Moslem et al., 2023). However, scant attention has been paid to the way in which translators, especially trainee translators, integrate LLMs into their daily workflows in practical terms. So far, Sahari et al. (2023) have conducted a cross-sectional study exploring attitudes of translation teachers and language-related major students towards ChatGPT and Google Translate, and the advantages and challenges brought by ChatGPT. The results show that among four language-related majors, all translation students prefer Google Translate over ChatGPT. Another study conducted by Zhang et al. (2025) investigates how translation students understand the benefits and challenges of using GenAI into their translation practice. While the study examined the functions of GenAI tools used by students in their translation practices, such as looking for background information, generating machine translation outputs, polishing human translations, and providing references for terminologies, its primary purpose was to explore trainee translators' perceptions of using GenAI in translation. However, the actual integration of LLMs into trainee translators' learning and practice remains underexplored. Therefore, this paper addresses this gap by investigating the use of LLMs across the three phases of translation services and their broader impact on human-AI communication with a focus on trainee translators.

To this end, the study examines when, in what contexts, and for what purposes trainee translators incorporate LLMs into their workflows, assessing their effectiveness and efficiency in different translation-related scenarios from a user-centered perspective. The study aims to identify the scenarios in which LLMs are most suitable and effective in students' translation workflow through a survey-based study. The results will serve as the initial step toward developing a large project: the design of an LLM translator guide to help trainee translators choose the most suitable LLM from among numerous options, including scenario-specific LLMs trained for different translation tasks and equipped with preset prompts. With its emphasis on translators in training, this research also seeks to contribute to the development of educational programs to better prepare future professionals for an AI-driven translation industry.

2 Literature Review

Apart from the technically oriented research mentioned above, current scholarly work in translation studies focusing on translators and users mainly addresses the perception and reception of new technologies, particularly AI, by translators (Wang et al., 2024; Wang & Zhang, 2024) and their impact on the language services industry

(Moorkens & Arenas, 2024; Shormani, 2024). More recently, a growing body of literature has begun to examine ethical concerns and sociotechnical effects associated with these innovations (Martinez Carrasco et al., 2024; Moorkens et al., 2024; O'Brien, 2024; Yu & Guo, 2024).

The pedagogical applications and implications of GenAI have also begun to attract considerable attention, particularly in the context of computer-assisted translator training (Ghosh & Chatterjee, 2024; Venkatesan, 2023). For example, Pym and Yu (2024) discuss the way in which translation technologies, including GenAI, can be integrated into language learning and translator training. Similarly, Peng et al. (2024) dedicates an entire section to pedagogy, including insights into students' experiences with, and feedback on, the use of translation technology.

Nevertheless, research on human-AI interaction and the comprehensive application of LLMs throughout all three phases of translation services—pre-production, production, and post-production—remains limited. While certain studies have examined prompting LLMs for translation tasks (Pourkamali & Ebrahim Sharifi, 2024; Zhang et al., 2023), the potential of LLMs to support functions beyond linguistic transfer through prompt engineering has received but little attention. Yamada (2023) investigates ChatGPT's customizability, for instance, but limits its analysis to prompt engineering for enhancing translation quality, so that further research is needed to examine its broader applications within a translator's workstation.

3 LLM-Activated Scenarios

3.1 Constitution of a translator's workstation

Since the concept of the translator's workstation emerged in the 1960s, numerous scholars, beginning with Martin Kay (1980), have attempted to define the range of facilities it might encompass. Among the key contributions to this discourse, Melby (1992) identifies three levels of functions for a translator workstation: (1) word processing, telecommunications and terminology management; (2) text analysis, dictionary lookup, and bilingual text retrieval; and (3) an interface to machine translation systems (147). More recently, Alonso and Nunes Vieira (2017) have updated Kay's (1980) seminal idea of a translator's amanuensis by proposing the Translator's Amanuensis 2020, which serves both "the general public in their daily translating needs, providing instant machine translation (henceforth referred to as 'the utility level'), and different actors involved with translation in professional settings" (349). Specifically, TA2020 (Alonso & Nunes Vieira, 2017) incorporates the following abilities:

> a) parse the source content (whether written or audio-visual);
> b) identify keywords (key concepts), topics, and genre;
> c) mine virtual content (publicly available and private knowledge bases) and social media in order to find relevant and reliable sources of information to be consulted in the translating process (websites, parallel multilingual content, images, augmented reality output, videos, news, reports), previous translations, and relevant multimodal content. (351)

Ideally, as a critical component of a modern translator's workstation, LLMs should be capable of performing many of these functions while addressing both source- and target-language perspectives.

3.2 Formulation of LLM-activated scenarios

In the preliminary stage of the study, we hypothesized that it was the "chatbot" function of LLMs that would be active when performing translation tasks, particularly their multi-turn dialogue capabilities (Bang et al., 2023). Translation is a decision-making process involving "a series of a certain number of consecutive situations imposing on the translator the necessity of choosing among a certain (and very often exactly definable) number of alternatives," as Levý points out (1967, p. 1171). In this sense, whenever a translator needs to come to a decision, LLMs can provide contextually relevant suggestions, thereby greatly expanding the scope of its application and utility.

Moreover, the real-time interactive query function allows LLMs to answer questions, resembling the search/query function of an internet browser. This means that whenever a translator seeks information, he or she would be able to apply directly to an LLM for assistance. These information retrieval and feedback-seeking functions are the most important ones throughout the process.

To better identify the specific steps or scenarios involved in translation, we drew on the states and events framework proposed by Hlebec (1989) as a reference and adapted it for the purpose of this study.

1. (Activating) knowledge required for an interpretation of the original
2. Choosing the code
3. Interpreting the original
4. Deciding on, or recognizing the presence or absence of the original
5. Considering the form of the translation code
6. Deciding the degree of literalness
7. Determining the intentions
8. Deciding on the manner of conveying the original intentions
9. (Activating) the knowledge required for recoding

In addition, we included the following necessary tasks—checking and revision—as highlighted in ISO 17100:2015, which a translation service might require before the submission and release of a translation, as well as feedback after submission:

10. Checking the target content for semantic, grammatical, and spelling issues, as well as omissions and other errors
11. Examining the target language content against the source language content for any errors, for suitability purpose, and for making corrections
12. Client feedback and satisfaction assessment

To prepare the design of our survey study, we further elaborated on these 12 scenarios and concretized them with a detailed list of functions for LLMs, inspired by Siu (2023):

1. Providing summaries of source texts
2. Highlighting key terms or phrases that require special attention
3. Offering background knowledge or explanations for culturally specific references
4. Suggesting appropriate translations for domain-specific terms
5. Retrieving definitions and usage examples from bilingual corpora or glossaries
6. Automatically identifying inconsistencies in terminology across the text
7. Deciding between literal and free translation based on the purpose of the text
8. Choosing appropriate style, tone, and register for the target audience
9. Resolving ambiguities in the source text
10. Answering specific questions on terminology, grammar, or cultural references
11. Providing links to relevant external resources
12. Acting as an advanced search engine
13. Identifying and correcting grammatical, semantic, or stylistic issues in the target text
14. Comparing the translation with the source text to ensure fidelity and alignment
15. Assessing the target text's suitability for its intended purpose and audience
16. Simulating a client to provide feedback
17. Analyzing client feedback to identify recurring issues or preferences
18. Providing suggestions for future

Phase	**Scenarios**
Pre-production	Summarizing the content of the source text
	Highlighting key terms or phrases that require special attention
	Providing background knowledge or external resources for understanding the source text
Production	Answering specific questions about terminology, grammar, or cultural references
	Suggesting appropriate style, tone, and register for the translation
	Providing translation references for sentences or paragraphs
	Identifying (and correcting) grammatical, semantic, or stylistic issues in the target text
	Examining whether the translation meets the standard of classic translation norms like "faithfulness, expressiveness and elegance"
Post-production	Providing feedback from the target audience's perspective
	Providing suggestions for future translations based on past feedback

Table 2: Ten Scenarios where Trainee Translators might Use LLMs Throughout the Translation Process

translations based on past feedback

4 Methodology

4.1 Design of the survey

A questionnaire was designed for the purposes of this study in order to investigate the way in which trainee translators use LLMs during a translation task (a task serving the same function for trainee translators as a translation service does for profession translators) by examining three aspects: the frequency of using LLMs in different scenarios, the prompting times in each scenario, and their satisfaction with the performance of LLMs in these scenarios (see Appendix A). The frequency of their LLM usage is used to identify situations where trainee translators commonly use LLMs during the translation process. The prompting times are expected to indicate the extent to which trainee translators strive to interact with LLMs and the efficiency of LLMs when used for different purposes, as fewer rounds of interaction improve user experience. In this context, prompting times refer to the average number of prompts given to the LLM to achieve a specific goal. For example, a trainee translator may prompt an LLM five times to search the background information on a culture-specific term or prompt an LLM three times to check the accuracy of a translation. The effectiveness and suitability of current LLMs under different circumstances is surveyed in "translators' satisfaction with LLMs".

The 18 scenarios introduced in Section 3.2, which aim to cover every possible situation where translators might resort to LLMs for a translation task, were further categorized into pre-production, production, and post-production scenarios, based on the phases and stages described in Table 1. Some overlapping scenarios have been streamlined and modified to ensure clearer distinctions and enhance the understanding of participants. As a result, we produced a table of ten refined scenarios (see Table 2).

In the research for the questionnaire, for each scenario, participants were first asked to specify the frequency of their LLM usage and were provided with four options: "never or rarely", "sometimes", "often", and "always". When participants chose the latter of the three options, which implied that they had access to LLMs and used them for a certain purpose, they would be further asked about their interaction times with LLMs on average and to rate their performance. However, for those who had "never or rarely" used LLMs in a certain scenario, the questionnaire offered options which were carefully designed to capture the possible reasons, including "I have never thought of using LLMs in this way", indicating their lack of understanding of

Phase	Scenarios	Metrics		
		Frequency Score	Prompting Times	Satisfaction Score
Pre-production	Summarizing the content of the source text	5	9	8
	Highlighting key terms or phrases that require special attention	7	10	5
	Providing background knowledge or external resources for understanding the source text	2	7	2
Production	Answering specific questions about terminology, grammar, or cultural references	1	8	1
	Suggesting appropriate style, tone, and register for the translation	6	2	6
	Providing translation references for sentences or paragraphs	4	6	10
	Identifying (and correcting) grammatical, semantic, or stylistic issues in the target text	3	4	4
	Examining whether the translation meets the standard of classic translation norms like "faithfulness, expressiveness and elegance"	8	1	9
Post-production	Providing feedback from the target audience's perspective	10	5	3
	Providing suggestions for future translations based on past feedback	9	3	7

Table 3: Rankings of the Ten Scenarios Based on the Three Metrics

the use of LLMs or awareness of this step during their translation practice, "I think LLMs' answers are not reliable", implying their distrust of LLMs, and "I think LLMs' answers are not useful", denoting the technical limitations of LLMs. If none of the options was suitable, participants were also asked to write down other underlying reasons. In addition, at the end of the questionnaire, participants were invited to list their most commonly used LLMs.

4.2 Participants

In this study, all the participants were first-year MA students enrolled in the Translation and Interpreting Studies program (with either a Translation and Interpreting major or a Translation plus New Technologies major) or the Simultaneous Interpreting program at the university where the researchers of this paper currently work. All participants were native Chinese speakers whose working language pair was Chinese and English, with IELTS scores of at least 7, who had completed at least one translation-related course during their postgraduate studies and had experience working on translation tasks both individually and in groups. All participants had been introduced to LLMs by their instructors, were familiar with LLMs, and had prior experience of using them in their translation.

4.3 Procedures

The questionnaire content was first submitted to the Applied Psychology Institutional Review Board of the university for an ethical check. Following approval, questionnaires with detailed instructions were distributed to participants via WJX.CN[1], an electronic survey platform widely recognized in China. A total of 50 questionnaires were collected, of which 41 were deemed valid for research purposes.

4.4 Data processing

To investigate the using frequency of each scenario, the interactions with LLMs, and the participants' evaluation of the performance of LLMs, the study employed a weighted average approach to calculate three metrics: frequency scores, prompting times, and satisfaction scores. For the frequency scores, participants' responses were weighted as follows: 0 point for "Never or Rarely," 1 point for "Sometimes," 2 points for "Often," and 3 points for "Always." To rank the prompting times, we assigned 1 point for the option "1-5," 2 points for "6-10," 3 points for "11-15," 4 points for "16-20," and 5 points for "Over 20." For the satisfaction scores, participants rated their satisfaction on a 5-point Likert scale (with 5 representing the highest level of LLM performance). Weighted averages were calculated for all three metrics across the ten scenarios (see Appendix B), and the scenarios were ranked in descending order to identify the most frequently used scenarios, the highest prompting times, and the highest satisfaction scores. Table 3 exhibits the rankings of the ten scenarios for the three metrics.

To better understand the performance of LLMs in each scenario, the researchers calculated the average number of prompts and average satisfaction score of the ten scenarios, then compared the prompting times and satisfaction score of each scenario with the corresponding averages. If the prompting times of a scenario was higher than the average, it may suggest that more time and energy were invested in these scenarios, indicating low efficiency in LLM performance. Conversely, if the prompting times of a scenario was lower than the average, it could mean less efforts spent on that scenario and more efficient LLM performance. Similarly, if the satisfaction score of a scenario was higher than the average, it may suggest participants' satisfaction with LLMs' performance in this scenario. However, if the satisfaction score of a scenario was lower than the average, it could mean the unsatisfactory performance of LLMs in this scenario.

5 Analysis

5.1 The interrelationship among the metrics

Given the primary goal of the study—to explore trainee translators' use of LLMs in their translation workflow, the analysis started from categorizing scenarios into two types, those where translation students regularly used LLMs and those where they rarely did, based on the ranking of the frequency scores. Then, the study examined the prompting times and satisfaction scores of each scenario to understand their popularity, as these two metrics respectively reflected the efficiency and effectiveness of LLM use. For instance, a high satisfaction score of a scenario may explain the

[1] https://www.wjx.cn/

frequent use of LLMs under this circumstance, while a high prompting count may suggest more effort was required to prompt the LLM to achieve the goal in this scenario and thus indicate a less satisfactory evaluation and less frequent use.

5.2 Seven regular scenarios where trainee translators use LLMs

In the ranking of the ten scenarios based on the frequency scores (see Table 3, Column Frequency Score), the top seven were recognized by over 50 percent of participants as regular scenarios where they used LLMs (i.e., participants selected "Sometimes," "Often," or "Always" as their response). It should be noted that all seven scenarios belonged to pre-production and production stages, indicating that current LLMs were generally more suitable and useful in these phases from the perspective of trainee translators. For these seven regular scenarios, four distinct roles played by LLMs can be observed (see Table 4). In other words, LLMs, like trainee translators' assistants, are capable of taking on the four specific roles in their translation workflows.

Roles	Functions
Corrector	To proofread trainee translators' work and to identify grammatical, semantic, or stylistic issues
Explainer	To explain various aspects for trainee translators, including terminology, background knowledge, and register of the source text
Generator	To generate new content by offering translations for certain sentences or paragraphs
Summarizer	To read information, extract key points and summarize the content. Examples include highlighting critical parts that need special attention during translation or summarizing the content of the source text

Table 4: LLMs' Four Roles in Translation Tasks

The comparison between satisfaction score on average and that of each scenario showed that, four out of the seven regular scenarios—shaded in Table 3, Column Satisfaction Score—namely "answering specific questions about terminology, grammar, or cultural references", "providing background knowledge or external resources for understanding the source text", "identifying (and correcting) grammatical, semantic, or stylistic issues in the target text", and "highlighting key terms or phrases that require special attention", scored above average, which, aligns with their frequent use by trainee translators and, to some extent, explains why these functions were frequently used by trainee translators. Trainee translators were satisfied with LLMs' performance in these scenarios, which belong to the three roles—Corrector, Explainer, and Summarizer. However, though the researcher had assumed that students' low level of satisfaction with a certain scenario should be reflected in a less frequent use, the remaining three regular scenarios scored below the average satisfaction score, indicating that some regular scenarios are particularly unsatisfactory for the trainee translators. Notably, the scenario "providing translation references for sentences or paragraphs" ranked fourth in frequency of use but last in satisfaction, suggesting that while the trainee translators had a strong demand for machine translation in their work, current LLM-based machine translation failed to meet their requirements, an issue that warrants further investigation.

In addition, the comparison between the average prompts and the prompting times of each scenario demonstrated that of the seven scenarios, five scenarios—shaded in Table 3, Column Prompting Times—including "summarizing the content of the source text", "highlighting key terms or phrases that require special attention", "providing background knowledge or external resources for understanding the source text", "answering specific questions about terminology, grammar, or cultural references", and "providing translation references for sentences or paragraphs" ranked below the overall average level. Considering their satisfaction scores, the less prompting times in scenarios including "highlighting key terms or phrases that require special attention", "providing background knowledge or external resources for understanding the source text", and "answering specific questions about terminology, grammar, or cultural references" suggest a high efficiency of LLMs' performance, explaining why trainee translators have demonstrated strong satisfaction with the three scenarios. However, "summarizing the content of the source text" and "providing translation references for sentences or paragraphs"

scenarios, though requiring fewer prompts compared to the average, scored lower than the average satisfaction level. This suggests that it is possible students may have obtained outputs from LLMs that were far from satisfactory ones and thus gave up prompting after first several rounds of interactions. As for the rest two scenarios ranking above the average prompts, trainee translators' satisfaction with "identifying (and correcting) grammatical, semantic, or stylistic issues in the target text" was higher than the average level, indicating that the students had a great need for explanations on the above issues, and that this scenario was of great importance to their translation practice. In contrast, "suggesting appropriate tone, style, or register for the translation" scored below the average satisfaction level, implying that students failed to obtain satisfactory answers after multiple turns of prompts. These findings demonstrate the need to develop prompts tailored to specific tasks, with the aim of maximizing the effectiveness of the initial response.

5.3 Three scenarios where trainee translators rarely use LLMs

Meanwhile, more than 60 percent of participants reported that they had never, or rarely, asked LLMs to "examine whether the translation meets the standard of classic translation norms like 'faithfulness, expressiveness and elegance'", "provide suggestions for future translations based on past feedback", or "provide feedback from the target audience's perspective".

When asked for the reasons, over 80 percent of participants stated that they had never thought of using LLMs to "provide suggestions for future translations based on past feedback", or to "provide feedback from the target audience's perspective", both of which belong to the post-production stage. One possible explanation is that trainee translators or translation training programs do not attach great importance to this stage, despite its importance in improving the quality of a final translation product and trainee translators' competence in the long run by providing continuous feedback and suggestions to support their development. Another possibility is that trainee translators believe post-production jobs should be performed by human beings rather than LLMs and have therefore never tried to use LLMs for this stage. However, it is worth noting that the satisfaction score for the performance of LLMs in "providing feedback from the target audience's perspective" ranked 3rd across the ten scenarios. To some extent, this suggests that LLMs are effective and useful for those who regularly use them in this scenario, proving the suitability of LLMs at the post-production stage. Raising awareness of these benefits could promote the use of LLMs in post-production among trainee translators. In addition, over 50 percent of respondents claimed that they had never thought of asking LLMs to "examine whether the translation meets the standard of classic translation norms like 'faithfulness, expressiveness and elegance'". The results indicate that trainee translators tend to pay less attention to translation norms during the production stage. One avenue for further development could be to incorporate translation norms into the design for prompt engineering, in addition to calling on translator trainers to encourage a combination of theory and practice in teaching AI-enhanced translation activities.

The satisfaction scores for two of the three scenarios— "examining whether the translation meets the standard of classic translation norms like "faithfulness, expressiveness and elegance" and "providing suggestions for future translations based on past feedback"—were relatively low. This could be attributed to poor human-AI communication and/or the current technical limitations of LLMs, as evidenced by the ranking of the prompting times, where these two scenarios were ranked among the top three. Although detailed reasons have not yet been explored, it is probable that translators may have to invest much more effort when interacting with LLMs in these situations. These results suggest the need to improve trainee translators' prompt engineering skills and fine tune LLMs to meet user expectations.

5.4 Trainee translators' commonly used LLMs

The list of LLMs used by participants, along with the frequency of their mentions in the questionnaire, is presented in Table 5.

As shown in the table, trainee translators tend to prefer open-source, general-purpose LLMs over translation-specific LLMs or LLM-integrated CAT tools. It should be noted that one student mentioned DeepL—which is not an

LLM—in the survey, indicating that there are still students preferring traditional machine translation tools rather than LLMs. One possible explanation could be students' limited access to commercial models designed specifically for translation tasks. In addition, these commercial models are often tailored to the needs of professional translators and thus may not fulfill the expectations or learning needs of students.

Name	Times
ChatGPT	37
Kimi	14
Cici (Doubao)	8
ERNIE Bot (Wenxinyiyan)	3
Claude	2
Deepseek	2
Tongyi Qianwen	2
Gemini	1
Grammarly	1
WPS Lingxi	1
Quark	1

Table 5: The LLMs Mentioned by Participants and the Number of Times they were Mentioned

Of the listed LLMs, ChatGPT was the most popular one. One contributing factor may have been its advanced intelligence. After GPT-4 was launched, it was tested to solve problems in various cases more effectively than the original ChatGPT and to perform tasks at a level comparable to that of human beings (Bubeck et al., 2023). Such results have boosted its reputation and popularity. Another possible reason is that, as an LLM targeting global users, ChatGPT performs better in generating English texts compared with Chinese domestic LLMs. The next two LLMs, Kimi and Cici, are both developed in China. Although they were less widely favored than ChatGPT, the frequency with which they were mentioned by participants might indicate a growing preference among trainee translators for domestic LLMs. The researcher assumes that users might be satisfied with their performance in understanding Chinese text due to the possibility that their training data is more closely aligned with Chinese culture. However, so far, no relevant studies have confirmed this assumption, nor is there any evidence on the sources of the training data used by Kimi and Cici.

5.5 Insights for the design of an LLM translator guide

The results of the questionnaire provide insights into the development of an LLM translator guide—a chatbot designed for translators' workstations, which aggregates scenario-specific LLMs and serves as a reference tool to direct translators to the appropriate LLM for different scenarios and provides ready-to-use prompts.

The findings indicate that trainee translators rely more on LLMs during the pre-production and production stages and engage with these tools less frequently in the post-production phase. Nevertheless, an effective LLM translator guide could cover scenarios across all three stages given the potential of LLMs to assist trainee translators throughout their workflow. Therefore, a practical starting point from which to develop the initial version of the LLM translator guide would be to focus on all scenarios identified and analyzed in this study, with particular attention to the post-production stage.

The results also shed light on the design of scenario-specific LLMs. On the one hand, for scenarios where current LLMs have already met basic requirements, scenario-specific LLMs could build on popular tools such as ChatGPT and Kimi, focusing on fine-tuning the models as well as improving the design of user interfaces and, ultimately, user experience. On the other hand, for scenarios where current LLMs have so far failed to meet the needs of trainee translators, the challenge is not only to design scenario-specific LLMs but also to ensure that the LLM translator guide optimizes existing functions by incorporating guidance and examples for prompt engineering. For instance, in scenarios where trainee translators currently experience more rounds of interaction compared with less effort-cost scenarios, priority should be given to improving the prompt engineering skills of trainee translators and the ability of the LLM translator guide to interpret prompts and provide more targeted responses, which would enhance the efficiency and effectiveness of human-AI communication.

Furthermore, when designing the LLM translator guide, clear instructions explaining its functions should be included to prevent its underuse due to insufficient awareness of specific features, as seen in certain post-production scenarios. These instructions need to be

accompanied by training to ensure students understand how to use the translator guide across the pre-production, production, and post-production stages. Such preparation would not only enhance the potential of the LLM translator guide as an innovative teaching and learning tool but also better prepare trainee translators to enter the AI-integrated translation workflow and industry in the future.

5.6 Limitations and future research plan

This study has identified regular scenarios where trainee translators commonly use LLMs or rarely resort to LLMs and has surveyed their regularly utilized LLMs when carrying out their translation tasks. However, several unresolved issues remain. The first issue is exploring why LLMs including ChatGPT, Kimi, and Cici were the most popular choices among the trainee translators. It is also worth exploring why the trainee translators tended not to use LLMs at the post-production stage—whether it was because of their lack of consideration for this stage or limitations in the ability of LLMs to perform tasks effectively. There has also been insufficient investigation into trainee translators' evaluations of the performance of LLMs. While they assessed the overall performance of LLMs, the satisfactoriness of their specific functions or design in certain scenarios, as well as the disadvantages that need improvement, are still unknown. A more comprehensive understanding is therefore needed to facilitate the design of scenario-specific LLMs in the future.

Therefore, in the follow-up research, we intend to conduct focus group interviews to explore translators' use of LLMs, their evaluations, and suggestions for LLM improvement. In terms of their use of LLMs, they will first be asked to share the reasons why they prefer to use certain LLMs during their translation practice. Their answers will help address the first issue mentioned above. In this part, they will also be asked to describe how they use LLMs across the three stages, which will inform the researchers of the details concerning their interactions with LLMs and help explain why they rarely use them during the post-production stage. As for their evaluation of LLMs and suggestions for improvement, the researchers will invite participants to systematically evaluate their performance, identify the deficiencies of current LLMs, and share their opinions on how these tools can be improved to meet their needs. Participants' answers to these two aspects will further clarify their rating of LLM performance.

Furthermore, since this study has identified the regular roles played by LLMs and the scenarios where trainee translators might use them, and given that the LLM translator guide is intended not only to serve as a specialized tool for trainee translators but also to support their translation learning, the researchers also aim to design a framework in the future to evaluate the performance of LLMs specifically for translation education from a user perspective. Drawing on existing evaluation frameworks based on user experience, the researchers are currently developing a customized framework to evaluate the use of LLMs in translation classrooms, which, broadly speaking, will be conducive to providing an AI-driven, immersive learning experience for trainee translators as well as promoting the integration of AI into translation pedagogy.

6 Conclusion

As a preliminary step towards building an LLM translator guide for trainee translators, this study has investigated the scenarios in which trainee translators rely on LLMs during their translation workflow, based on questionnaire research. The findings revealed that interactions between trainee translators and LLMs occurred mainly in the pre-production and production stages, where LLMs were used for tasks such as question answering, correction, content generation, and summarization. In contrast, the post-production stage saw less engagement with LLMs. Moreover, despite the fact that trainee translators have already started to integrate LLMs into their translation workflow, their evaluation of the performance of LLMs revealed areas for improvement.

This study has laid the groundwork for further research and development to optimize human-AI collaboration in translation. It is hoped that the final product—the LLM translator guide—will enhance the competence of trainee translators and better prepare them for human-AI collaboration in future practice. If it works, the project will be extended to design a guide for professional translators as well, to improve their work efficiency and enable them to thrive in an AI-integrated translation industry.

References

Elisa Alonso, & Lucas Nunes Vieira. 2017. The Translator's Amanuensis 2020. *JoSTrans: The Journal of Specialised Translation*(28):345-361. http://www.jostrans.org/issue28/art_alonso.php

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, & Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity.arXiv:2302.04023. Retrieved February 01, 2023, from https://ui.adsabs.harvard.edu/abs/2023arXiv230204023B.

Rachel Bawden, & François Yvon. Year. Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM.In *Proceeding Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM*, European Association for Machine Translation, pages 157-170. https://aclanthology.org/2023.eamt-1.16/

Santosh Bhupathi. 2025. Role of Databases in GenAI Applications. *arXiv preprint arXiv:2503.04847*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, & Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4.arXiv:2303.12712. Retrieved March 01, 2023, from https://ui.adsabs.harvard.edu/abs/2023arXiv230312712B.

Sourojit Ghosh, & Srishti Chatterjee. 2024. Machine Translation, Large Language Models, and Generative AI in the University Classroom:Toward a Pedagogy of Care. In E. Monzó-Nebot & V. Tasa-Fuster (ed.), *The Social Impact of Automating Translation*, pages. Routledge

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, & Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation.arXiv:2302.09210. Retrieved February 01, 2023, from https://ui.adsabs.harvard.edu/abs/2023arXiv230209210H.

Martin Kay. 1980. The Proper Place of Men and Machines in Language Translation. https://aclanthology.org/www.mt-archive.info/70/Kay-1980.pdf

Jiři Levý. 1967. Translation as A Decision Process. In (ed.), *To honor Roman Jakobson : essays on the occasion of his 70. birthday, 11. October 1966*, pages 1171-1182. De Gruyter Mouton.https://doi.org/doi:10.1515/9783111349121-031.

Robert Martinez Carrasco, Anabel Borja Albi, & Łucja Biel. 2024. Legal translation in the face of (de)globalisation. The impact of human development, polycrises and technological disruptions in language service provision. *MonTI. Monographs in Translation and Interpreting*(16). https://www.e-revistes.uji.es/index.php/monti/article/view/8116

Alan Melby. 1992. The translator workstation. In J. Newton (ed.), *Computers in Translation*, pages 147-165. Routledge.https://doi.org/10.4324/9780203128978.

Shereen A Mohamed, Ashraf A Elsayed, YF Hassan, & Mohamed A Abdou. 2021. Neural machine translation: past, present, and future. *Neural Computing and Applications*, *33*:15919-15931.

Joss Moorkens, & Ana Guerberof Arenas. 2024. Artificial intelligence, automation and the language industry. In M. Gary, E.-D. Maureen, & A. Erik (ed.), *Handbook of the Language Industry*, pages 71-98. De Gruyter Mouton.https://doi.org/doi:10.1515/9783110716047-005.

Joss Moorkens, Sheila Castilho, Federico Gaspari, Antonio Toral, & Maja Popović. 2024. Proposal for a Triple Bottom Line for Translation Automation and Sustainability: An Editorial Position Paper. *The Journal of Specialised Translation*(41):2-25. https://doi.org/10.26034/cm.jostrans.2024.4706.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, & Andy Way. Year. Adaptive Machine Translation with Large Language Models.In *Proceeding Adaptive Machine Translation with Large Language Models*, European Association for Machine Translation, pages 227–237. https://aclanthology.org/2023.eamt-1.22/

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, & Ajmal Mian. 2023. A Comprehensive Overview of Large Language Models.arXiv:2307.06435. Retrieved July 01, 2023, from https://ui.adsabs.harvard.edu/abs/2023arXiv230706435N.

Sharon O'Brien. 2024. Human-Centered augmented translation: against antagonistic dualisms. *Perspectives*, *32*(3):391-406. https://doi.org/10.1080/0907676X.2023.2247423.

Yuhong Peng, Huihui Huang, & Defeng Li. 2024. *New Advances in Translation Technology: Applications and Pedagogy*. Springer. https://doi.org/10.1007/978-981-97-2958-6.

Nooshin Pourkamali, & Shler Ebrahim Sharifi. 2024. Machine Translation with Large Language Models: Prompt Engineering for Persian, English, and Russian Directions.arXiv:2401.08429. Retrieved January 01, 2024, from https://ui.adsabs.harvard.edu/abs/2024arXiv240108429P.

Anthony Pym, & Hao Yu. 2024. *How to Augment Language Skills*. Routledge. https://doi.org/10.4324/9781032648033.

Yousef Sahari, Abdu M Talib Al-Kadi, & Jamal Kaid Mohammed Ali. 2023. A cross sectional study of ChatGPT in translation: Magnitude of use, attitudes, and uncertainties. *Journal of Psycholinguistic Research*, *52*(6):2937-2954.

Mohammed Q. Shormani. 2024. Artificial intelligence contribution to translation industry: looking back and forward.arXiv:2411.19855. Retrieved November 01, 2024, from https://ui.adsabs.harvard.edu/abs/2024arXiv241119855S.

Sai Cheong Siu. 2023. ChatGPT and GPT-4 for Professional Translators: Exploring the Potential of Large Language Models in Translation.

International Organization for Standardization. 2015. Translation services — Requirements for translation services of the Standard. ISO 17100:2015, International Organization for Standardization.

International Organization for Standardization. 2024. Translation projects — General guidance of the Standard. ISO 11669:2024, International Organization for Standardization.

Hari Venkatesan. 2023. Technology preparedness and translator training. *Babel*, *69*(5):666-703. https://doi.org/https://doi.org/10.1075/babel.00335.ven.

Lulu Wang, Simin Xu, & Kanglong Liu. 2024. Understanding Students' Acceptance of ChatGPT as a Translation Tool: A UTAUT Model Analysis.arXiv:2406.06254. Retrieved June 01, 2024, from https://ui.adsabs.harvard.edu/abs/2024arXiv240606254W.

Yun Wang, & Zheng Zhang. 2024. The Pitfall and Relief of ChatGPT Artificial Intelligence Translation. *Chinese Translators Journal*, *45*(2):95–102.

Wordscope. *Translate your documents faster and better thanks to Artificial Intelligence*. https://pro.wordscope.com/

Masaru Yamada. Year. Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability.In *Proceeding Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability*, Asia-Pacific Association for Machine Translation, pages 195-204. https://aclanthology.org/2023.mtsummit-users.19/

Hao Yu, & Yunyun Guo. 2024. Risk and Transcendence: an ethical analysis of ChatGPT enabling translation. *Chinese Translators Journal*, *45*(4):115–122.

Biao Zhang, Barry Haddow, & Alexandra Birch. 2023. Prompting Large Language Model for Machine Translation: A Case Study.arXiv:2301.07069. Retrieved January 01, 2023, from https://ui.adsabs.harvard.edu/abs/2023arXiv230107069Z.

Wenkang Zhang, Albert W Li, & Chenze Wu. 2025. University students' perceptions of using generative AI in translation practices. *Instructional Science*:1-23. https://doi.org/https://doi.org/10.1007/s11251-025-09705-y.

Appendix A Questionnaire on Trainee Translators' Use of LLMs Throughout the Translation Process

1. I ask LLMs to summarize the content of the source text.
   - Never or Rarely
   - Sometimes
   - Often
   - Always

(If the participant chooses "never or rarely", he/she will be asked to answer the following question)

Following question: Please choose your reason (Multiple-select question).

- I have never thought of using LLMs in this way.
- I think LLMs' answers are not reliable.
- I think LLMs' answers are not useful.
- Other (Blank)

(If the participant chooses "sometimes" to "always", he/she will be asked to answer the following two questions)

Following question 1: On average, how many prompts (e.g., instructions, clarifications, or

follow-up requests) do you use when asking LLMs to summarize the content of the source text?

- 1-5
- 6-10
- 11-15
- 16-20
- Over 20

Following question 2: On a scale of 1 to 5, with 5 being the highest score, how would you rate the performance of LLMs in summarizing content?

- 1
- 2
- 3
- 4
- 5

2. I ask LLMs to highlight key terms or phrases that require special attention.

- Never or Rarely
- Sometimes
- Often
- Always

(If the participant chooses "never or rarely", he/she will be asked to answer the following question)

Following question: Please choose your reason (Multiple-select question).

- I have never thought of using LLMs in this way.
- I think LLMs' answers are not reliable.
- I think LLMs' answers are not useful.
- Other (Blank)

(If the participant chooses "sometimes" to "always", he/she will be asked to answer the following two questions)

Following question 1: On average, how many prompts (e.g., instructions, clarifications, or follow-up requests) do you use when asking LLMs to highlight key terms or phrases that require special attention?

- 1-5
- 6-10
- 11-15
- 16-20
- Over 20

Following question 2: On a scale of 1 to 5, with 5 being the highest score, how would you rate the performance of LLMs in highlighting key terms or phrases that require special attention?

- 1
- 2
- 3
- 4
- 5

3. I ask LLMs to provide background knowledge or external resources enabling me to understand the source text.

- Never or Rarely
- Sometimes
- Often
- Always

(If the participant chooses "never or rarely", he/she will be asked to answer the following question)

Following question: Please choose your reason (Multiple-select question).

- I have never thought of using LLMs in this way.
- I think LLMs' answers are not reliable.
- I think LLMs' answers are not useful.
- Other (Blank)

(If the participant chooses "sometimes" to "always", he/she will be asked to answer the following two questions)

Following question 1: On average, how many prompts (e.g., instructions, clarifications, or follow-up requests) do you use when asking LLMs to provide background knowledge or external resources for you to understand the source text?

- 1-5
- 6-10
- 11-15
- 16-20
- Over 20

Following question 2: On a scale of 1 to 5, with 5 being the highest score, how would you rate the performance of LLMs in providing background knowledge or external resources enabling you to understand the source text?

- 1
- 2
- 3
- 4
- 5

4. I ask LLMs to answer specific questions about terminology, grammar, or cultural references.

- Never or Rarely
- Sometimes
- Often
- Always

(If the participant chooses "never or rarely", he/she will be asked to answer the following question)

Following question: Please choose your reason (Multiple-select question).

- I have never thought of using LLMs in this way.
- I think LLMs' answers are not reliable.
- I think LLMs' answers are not useful.
- Other (Blank)

(If the participant chooses "sometimes" to "always", he/she will be asked to answer the following two questions)

Following question 1: On average, how many prompts (e.g., instructions, clarifications, or follow-up requests) do you use when asking LLMs to answer specific questions about terminology, grammar, or cultural references?

- 1-5
- 6-10
- 11-15
- 16-20
- Over 20

Following question 2: On a scale of 1 to 5, with 5 being the highest score, how would you rate the performance of LLMs in answering specific questions about terminology, grammar, or cultural references?

- 1
- 2
- 3
- 4
- 5

5. I ask LLMs to suggest appropriate style, tone, and register of the translation.

- Never or Rarely
- Sometimes
- Often
- Always

(If the participant chooses "never or rarely", he/she will be asked to answer the following question)

Following question: Please choose your reason (Multiple-select question).

- I have never thought of using LLMs in this way.
- I think LLMs' answers are not reliable.
- I think LLMs' answers are not useful.
- Other (Blank)

(If the participant chooses "sometimes" to "always", he/she will be asked to answer the following two questions)

Following question 1: On average, how many prompts (e.g., instructions, clarifications, or follow-up requests) do you use when asking LLMs to suggest appropriate style, tone, and register of the translation?

- 1-5
- 6-10
- 11-15
- 16-20
- Over 20

Following question 2: On a scale of 1 to 5, with 5 being the highest score, how would you rate the performance of LLMs in suggesting appropriate style, tone, and register of the translation?

- 1
- 2
- 3
- 4
- 5

6. I ask LLMs to provide translation references for sentences or paragraphs.

- Never or Rarely
- Sometimes
- Often
- Always

(If the participant chooses "never or rarely", he/she will be asked to answer the following question)

Following question: Please choose your reason (Multiple-select question).

- I have never thought of using LLMs in this way.
- I think LLMs' answers are not reliable.
- I think LLMs' answers are not useful.
- Other (Blank)

(If the participant chooses "sometimes" to "always", he/she will be asked to answer the following two questions)

Following question 1: On average, how many prompts (e.g., instructions, clarifications, or follow-up requests) do you use when asking LLMs to provide translation references for sentences or paragraphs?

- 1-5
- 6-10
- 11-15
- 16-20
- Over 20

Following question 2: On a scale of 1 to 5, with 5 being the highest score, how would you rate the performance of LLMs in providing translation references for sentences or paragraphs?

- 1
- 2
- 3

- 4
- 5

7. I ask LLMs to identify (and correct) grammatical, semantic, or stylistic issues in the target text.
- Never or Rarely
- Sometimes
- Often
- Always

(If the participant chooses "never or rarely", he/she will be asked to answer the following question)
Following question: Please choose your reason (Multiple-select question).
- I have never thought of using LLMs in this way.
- I think LLMs' answers are not reliable.
- I think LLMs' answers are not useful.
- Other (Blank)

(If the participant chooses "sometimes" to "always", he/she will be asked to answer the following two questions)
Following question 1: On average, how many prompts (e.g., instructions, clarifications, or follow-up requests) do you use when asking LLMs to identify and correct grammatical, semantic, or stylistic issues in the target text?
- 1-5
- 6-10
- 11-15
- 16-20
- Over 20

Following question 2: On a scale of 1 to 5, with 5 being the highest score, how would you rate the performance of LLMs in identifying and correcting grammatical, semantic, or stylistic issues in the target text?
- 1
- 2
- 3
- 4
- 5

8. I ask LLMs to examine whether the translation meets the standard of classic translation norms like "faithfulness, expressiveness and elegance".
- Never or Rarely
- Sometimes
- Often
- Always

(If the participant chooses "never or rarely", he/she will be asked to answer the following question)
Following question: Please choose your reason (Multiple-select question).
- I have never thought of using LLMs in this way.
- I think LLMs' answers are not reliable.
- I think LLMs' answers are not useful.
- Other (Blank)

(If the participant chooses "sometimes" to "always", he/she will be asked to answer the following two questions)
Following question 1: On average, how many prompts (e.g., instructions, clarifications, or follow-up requests) do you use when asking LLMs to examine whether the translation meets the standard of classic translation norms like "faithfulness, expressiveness and elegance"?
- 1-5
- 6-10
- 11-15
- 16-20
- Over 20

Following question 2: On a scale of 1 to 5, with 5 being the highest score, how would you rate the performance of LLMs in examining whether the translation meets the standard of classic translation norms like "faithfulness, expressiveness and elegance" ?
- 1
- 2
- 3
- 4
- 5

9. I ask LLMs to provide feedback from the target audience's perspective.
- Never or Rarely
- Sometimes
- Often
- Always

(If the participant chooses "never or rarely", he/she will be asked to answer the following question)
Following question: Please choose your reason (Multiple-select question).
- I have never thought of using LLMs in this way.
- I think LLMs' answers are not reliable.
- I think LLMs' answers are not useful.
- Other (Blank)

(If the participant chooses “sometimes” to “always”, he/she will be asked to answer the following two questions)
Following question 1: On average, how many prompts (e.g., instructions, clarifications, or follow-up requests) do you use when asking LLMs to provide feedback from the target audience’s perspective?
- o 1-5
- o 6-10
- o 11-15
- o 16-20
- o Over 20

Following question 2: On a scale of 1 to 5, with 5 being the highest score, how would you rate the performance of LLMs in providing feedback from the target audience’s perspective?
- o 1
- o 2
- o 3
- o 4
- o 5

10. I ask LLMs to provide suggestions for my future translations.
- o Never or Rarely
- o Sometimes
- o Often
- o Always

(If the participant chooses “never or rarely”, he/she will be asked to answer the following question)
Following question: Please choose your reason (Multiple-select question).
- o I have never thought of using LLMs in this way.
- o I think LLMs’ answers are not reliable.
- o I think LLMs’ answers are not useful.
- o Other <u>(Blank)</u>

(If the participant chooses “sometimes” to “always”, he/she will be asked to answer the following two questions)
Following question 1: On average, how many prompts (e.g., instructions, clarifications, or follow-up requests) do you use when asking LLMs to provide suggestions for your future translations?
- o 1-5
- o 6-10
- o 11-15
- o 16-20
- o Over 20

Following question 2: On a scale of 1 to 5, with 5 being the highest score, how would you rate the performance of LLMs in providing suggestions for your future translations?
- o 1
- o 2
- o 3
- o 4
- o 5

11. If there are other scenarios that are not mentioned above, please write them down.
<u>(Blank)</u>
12. Please write down the names of your most commonly used LLMs.
<u>(Blank)</u>

Appendix B Weighted Average Scores for Three Metrics

Scenarios	Weighted Average for Three Metrics		
	Frequency Score	Prompting Times	Satisfaction Score
Summarizing the content of the source text	1.10	1.36	3.25
Highlighting key terms or phrases that require special attention	0.76	1.32	3.41
Providing background knowledge or external resources for understanding the source text	1.51	1.53	3.64
Answering specific questions about terminology, grammar, or cultural references	1.71	1.46	3.77
Suggesting appropriate style, tone, and register for the translation	0.80	1.78	3.35
Providing translation references for sentences or paragraphs	1.22	1.53	3.13
Identifying (and correcting) grammatical, semantic, or stylistic issues in the target text	1.27	1.69	3.47
Examining whether the translation meets the standard of classic translation norms like “faithfulness, expressiveness and elegance”	0.56	1.93	3.14
Providing feedback from the target audience’s perspective	0.39	1.67	3.58
Providing suggestions for future translations based on past feedback	0.51	1.73	3.33

Introducing Quality Estimation to Machine Translation Post-editing Workflow: An Empirical Study on Its Usefulness

Siqi Liu
The Hong Kong Polytechnic University
si-qi.liu@connect.polyu.hk

Guangrong Dai
Guangdong University of Foreign Studies
carldy@163.com

Dechao Li*
The Hong Kong Polytechnic University
dechao.li@polyu.edu.hk

Abstract

This preliminary study investigates the usefulness of sentence-level Quality Estimation (QE) in English-Chinese Machine Translation Post-Editing (MTPE), focusing on its impact on post-editing speed and student translators' perceptions. It also explores the interaction effects between QE and MT quality, as well as between QE and translation expertise. The findings reveal that QE significantly reduces post-editing time. The examined interaction effects were not significant, suggesting that QE consistently improves MTPE efficiency across medium- and high-quality MT outputs and among student translators with varying levels of expertise. In addition to indicating potentially problematic segments, QE serves multiple functions in MTPE, such as validating translators' evaluations of MT quality and enabling them to double-check translation outputs. However, interview data suggest that inaccurate QE may hinder post-editing processes. This research provides new insights into the strengths and limitations of QE, facilitating its more effective integration into MTPE workflows to enhance translators' productivity.

1 Introduction

In a typical machine translation post-editing (MTPE) workflow, translators still need to spend a certain amount of time and effort on evaluating the quality of machine translation (MT) outputs to determine the cost-effectiveness of MTPE. To be more specific, if the MT output is of acceptable quality, post-editing is feasible; otherwise, translating from scratch may be more efficient. However, this process can be time-consuming, especially when the MT outputs are ultimately deemed unsuitable for post-editing. In order to achieve a quick turnaround, it is therefore necessary to speed up or even automate the process of evaluating whether MTPE is worthwhile (Alvarez-Vidal and Oliver, 2023).

The cost-effectiveness of MTPE can be evaluated from two different but interrelated perspectives: predicting the MT quality (Béchara et al., 2021; Specia et al., 2010), to see whether translators are going to work with "good" MT or "bad" MT, or predicting MTPE effort (Daems et al., 2017; Dai and Liu, 2024), to see how much effort, such as post-editing time and editing distance, is required by the PE task. In the field of computer science, both approaches are considered as Quality Estimation (QE)[1]. In contrast to traditional reference-based metrics such as BLEU (Papineni et al., 2002), QE estimates MT quality *without requiring reference translation* (Specia et al., 2010), making it particularly relevant for real-world translation scenarios. In the current research, we focus on the QE method that provides MT quality scores, rather than the one that estimates MTPE effort. The latter may be less straightforward for translators when making post-editing decisions, since a threshold that sets a point from which post-editing becomes translating from scratch (Do Carmo and Moorkens, 2020) has yet to be widely established.

The possible advantages of adopting QE to facilitate the MTPE workflow extend beyond streamlining the initial assessment of MTPE's cost-effectiveness. By providing information about the estimated quality of MT outputs, QE may help translators to allocate their efforts more effectively and focus on the outputs that deserve editing. On the one hand, they can spend minimal time on the

* Corresponding author

[1] However, since PE effort is a complex, multidimensional concept influenced by various factors — including but not limited to MT quality — and is not necessarily linearly related to MT quality (Alvarez-Vidal and Oliver, 2023; Krings, 2001), we argue for a clear distinction between the tasks of predicting MT quality and those of predicting PE effort, rather than grouping them into the same category.

Proceedings of Machine Translation Summit XX Volume 1, pages 485–495, June 23-27, 2025

least likely problematic MT outputs that require little intervention, avoiding making preferential edits. On the other hand, they can avoid wasting time reviewing and attempting to fix bad MT outputs unsuitable for post-editing (Moorkens et al., 2015; Specia et al., 2009). In addition, QE may free up time for translators to focus on tasks that are difficult to automate, such as creative translation. Translators themselves have also expressed the need for CAT tools to present MT quality information or to highlight problematic MT outputs requiring attention (Moorkens and O'Brien, 2013, 2017; Vieira and Alonso, 2018). While QE is not yet fully accurate in the realistic scenarios, its performance has steadily improved in recent years, particularly at the sentence-level (Blain et al., 2023; Specia et al., 2020, 2021; Zerva et al., 2022). Furthermore, the development of neural metrics alongside large language models offers the potential to further improve the accuracy and usability of QE (Zerva et al., 2024).

Despite the potential benefits and advancements of QE, it seems that QE has yet to be widely integrated in the real post-editing settings (Gilbert, 2022). One possible reason for this is the scarcity of CAT tools that can effectively incorporate QE. While some CAT tools, such as Trados, have recently started offering QE information, it is usually not freely accessible, posing a challenge to the widespread adoption of QE. Moreover, there is limited empirical evidence supporting the usefulness of QE in enhancing MTPE workflow, which makes it challenging for translators to embrace QE, as they may be uncertain about its practical value and impact on their work. In real-world applications, various factors could influence the effectiveness of QE, such as translators' attitudes towards QE, the accuracy of QE information, and the actual quality of MT outputs. Therefore, a critical question arises: to what extent and under what conditions can QE facilitate the MTPE processes?

In light of the above, this study investigates the usefulness of sentence-level QE[2] in the context of English-Chinese MTPE, taking both productivity and users' perceptions into consideration. It is expected that the current research can provide a more detailed understanding of QE's application in aiding post-editing tasks, shedding new lights on its strengths and limitations. This insight will contribute to a more effective integration of QE into the MTPE workflow, enhancing efficiency for translators. Specifically, it focuses on the following three research questions: 1. What is the impact of sentence-level QE on post-editing time? 2. Is the impact of sentence-level QE on post-editing time consistent across different conditions, in particular, varying levels of MT quality and translation expertise? 3. What are users' perceptions of the usefulness of sentence-level QE in post-editing?

[2]Based on the granularity of assessment, QE models can be classified into word, sentence, and document levels. This study specifically focuses on sentence-level QE.

2 Related Work

Existing research on the usefulness of sentence-level QE in the context of MTPE has primarily focused on its impact on MTPE productivity and translators' perceptions. While studies suggest that QE has the potential to enhance productivity, the evidence remains limited and mixed. For instance, Huang et al. (2014) observed a 10% productivity increase when QE information was provided during post-editing tasks. However, this improvement was measured against a human translation condition rather than a post-editing condition without QE. In other words, the productivity gains resulted from a combined effect of MT and QE, making it unclear how much QE alone directly contributed to the observed improvement. Similarly, Turchi et al. (2015) found a slight increase in post-editing speed, but this increase was not statistically significant. In Béchara et al.'s (2021) study, post-editing with QE resulted in lower average post-editing time, fewer keystrokes, and higher translation quality compared to the condition without QE. Despite these positive findings, the study did not report the statistical significance of these differences, which leaves the robustness of the improvements uncertain. Lee et al. (2021) explored QE within IntelliCAT, a CAT interface that provides three intelligent features, namely QE, translation suggestion, and word alignment. The results indicated a significant improvement in post-editing efficiency when working with IntelliCAT. However, as the tool incorporated both word-level and sentence-level QE alongside with other two features, it was difficult to determine the extent to which segment-level QE contributed to the increased post-editing speed.

In addition to productivity, it is essential to explore how translators perceive the usefulness of QE and the challenges they encounter when interacting with it. While earlier surveys revealed translators' interest in using QE information (Moorkens and

O'Brien, 2013, 2017; Vieira and Alonso, 2018), few studies have investigated the perceptions of translators after letting them actually work with QE during post-editing, and the findings have been inconclusive. Parra Escartín et al. (2017) collected translators' opinions on QE and revealed generally negative attitudes towards its usefulness. However, the reasons behind these negative results were not examined in this study. By contrast, most of the participants in Lee et al. (2021) expressed positive views on QE, particularly regarding its usefulness for proofreading purposes, such as double-checking the potential translation errors.

Apart from investigating the general impact of QE, efforts have also been made to consider additional factors and examine whether the usefulness of QE varies under specific conditions. For instance, given that QE was found to contribute only slight and insignificant global productivity gains in Turchi et al.'s (2015) study, the authors conducted an additional analysis to explore whether these marginal gains might become more pronounced under certain conditions. The analysis incorporated the length of source text (ST) and the quality of MT outputs, and the results suggested that QE led to significant productivity gains when the sentences were of medium length and had HTER[3] values between 0.2 and 0.5. The accuracy of QE has also been examined, with somewhat conflicting results. Parra Escartín et al. (2017) found that QE, especially good QE that provided a predicted quality score close to the actual score, significantly decreased post-editing time. However, Teixeira and O'Brien (2017) reported that no significant effect was introduced by QE, even when it was accurate. In addition, while not explicitly addressed as a variable of interest, Béchara et al. (2021) presented data pertaining to translation experience. In this study, despite varying levels of experience, all translators, with only one exception, increased their post-editing speed when QE information was provided.

In conclusion, there is a notable lack of empirical research on the effectiveness of presenting sentence-level QE information within the MTPE context, and the findings to date have been inconsistent. While considering additional factors has provided a more nuanced understanding of QE's impact on post-editing efficiency, more research is warranted. It should be noted that most previous studies have relied on basic statistical analyses, which may not fully capture the true impact of QE. Additionally, while professional translators have been the focus of these studies, the way student translators utilise QE information in the MTPE workflow has yet to be investigated, which can provide valuable insights into translation education.

3 Research Design and Methodology

3.1 Participants

Thirty-one first-year Master in Translation and Interpreting (MTI) students (6 males, 25 females) participated in the post-editing experiments. The average age of the participants was 23 years (range = 21-33, SD = 2.4). All students used Chinese as their L1 and English as L2, and have passed the Test for English Majors at Band4 (TEM4). Although they were in the same year of study, their translation expertise varied, as reflected by the levels of the China Accreditation Test for Translators and Interpreters (CATTI) [4]they had achieved. To be more specific, 23 participants had passed the CATTI Level 3 (Translator), while 8 had passed the CATTI Level 2 (Translator). However, none of them had worked as professional translators. While the participants had limited experience with MTPE, they generally held a positive attitude towards it, with an average rating of 6.16 (SD=0.86) on a seven-point scale, where '7' indicated a very positive attitude.

3.2 Materials

Given that this study focuses on the impact of QE on MTPE, it is essential to ensure the comparability between the materials used for the MTPE task without QE (Task 1) and the task with QE (Task 2). Specifically, textual characteristics, including ST complexity and MT quality, were controlled at a similar level across tasks, as suggested by previous research (Dai and Liu, 2024; Jia and Zheng, 2022). Each task [5] consisted of four short, self-contained news texts that required no specialist knowledge

[3]HTER (Human-targeted Translation Edit Rate) is a widely-used metric for assessing MT quality, which quantifies the number of edits necessary to transform the MT output into a good translation (Snover et al., 2006). It ranges from 0 to 1, with lower HTER representing higher MT quality.

[4]Recipients of CATTI Level 3 (translator) certificate are expected to complete general translation work, while those with a Level 2 certificate should be capable of handling complex translation tasks within a particular domain (http://www.catticenter.com/cattiksjj/1848)

[5]The materials, data, and script of statistical analyses used in the study are available at https://github.com/jam0127/QEresearch.

	Source Text				MT Output
	Word Count	Average Sentence Length	FRE	CAREC	MT Quality (mean/sd)
Task1 (without QE)	**304.00**	**13.62**	**63.32**	**0.14**	**2.58/0.10**
Text1	48.00	12.00	80.09	0.13	2.60/0.20
Text2	58.00	19.33	44.27	0.24	2.44/0.20
Text3	83.00	10.38	63.80	0.01	2.67/0.00
Text4	115.00	12.78	65.13	0.19	2.60/0.10
Task2 (with QE)	**295.00**	**12.90**	**59.20**	**0.11**	**2.56/0.19**
Text5	57.00	11.40	58.64	0.11	2.33/0.34
Text6	48.00	16.00	61.93	0.16	2.67/0.14
Text7	78.00	13.00	70.32	-0.01	2.75/0.13
Text8	112.00	11.20	45.91	0.19	2.48/0.07

Table 1: Summary of ST complexity and MT quality of the materials (a higher FRE score suggests lower complexity, while a higher CAREC score implies higher complexity)

for post-editing. As shown in Table 1, word count, average sentence length, and readability scores indicate that the two tasks were comparable in terms of ST complexity. Text readability was measured using two formulas: the Flesch Reading Ease (FRE) formula (Flesch, 1948), a traditional readability formula, and the Crowdsourced Algorithm of Reading Comprehension (CAREC) (Crossley et al., 2019), a newer formula. These two metrics focus on different aspects of text complexity: FRE relies on word length and sentence length, while CAREC is based on features pertaining to lexical sophistication and text cohesion. Therefore, the readability scores may vary when measured by different formulas. For this study, Task 1 received a higher FRE score than Task 2 on average, suggesting slightly lower complexity. However, according to CAREC, Task 1 was judged to be slightly harder to read. Despite these minor variations, the readability scores across both tasks were similar overall. Therefore, we concluded that the tasks were comparable in terms of readability.

The STs were translated by Baidu Translate, a mainstream NMT engine. Three second-year MA students in translation participated in the MT quality evaluation. They were not involved in the post-editing experiments. All of them had prior experience in annotating MT errors and had passed the CATTI Level 2 (Translator). The MT outputs were rated at the segment level using a three-point scale: a score of '1' suggested that the outputs require extensive editing or complete re-translation, while a score of '3' indicated minimal or no editing was needed. The inter-rater agreement was strong and significant (Kendall's W=0.705, p<0.05). Table 1 shows that the overall MT quality was comparable between the two tasks.

3.3 Research Procedures

To ensure the ecological validity of this study, we adopted YiCAT, a Chinese online CAT platform employed in the realistic translation scenario, along with its QE system. Since 2022, YiCAT has integrated QE as an optional feature within its interface, allowing translators to choose whether to display the information of estimated MT quality. Figure 1 and Figure 2 illustrate the interface used by participants for Task 1 and Task 2 respectively. The only difference between the two task interfaces lies in the third column (from left to right). In Task 1, it did not display any QE information (AT in this column is short for automatic translation). In Task 2, the column presented QE scores: "A" indicated that MT output is of good quality and requires minimal editing, "B" denoted medium-quality MT requiring moderate editing, and "C" represented poor-quality MT outputs that need extensive editing or retranslation. All editing actions were performed in the target text area (the second column from left to right).

The post-editing experiment was conducted on the campus of Guangdong University of Foreign Studies in October 2022. One day prior to the experiment, participants received a video tutorial on using YiCAT [6] and were required to complete a practice task to familiarise themselves with the plat-

[6]It is important to note that in YiCAT, the time spent on a segment would not be recorded if no edits were made to the segment. To ensure that the post-editing time for each segment was captured, participants were instructed to type '1' at the end of a segment if they believed the MT output required no editing. This additional step was also emphasised during the training session conducted before each post-editing experiment.

Figure 1: The YiCAT interface (Task 1, without QE information)

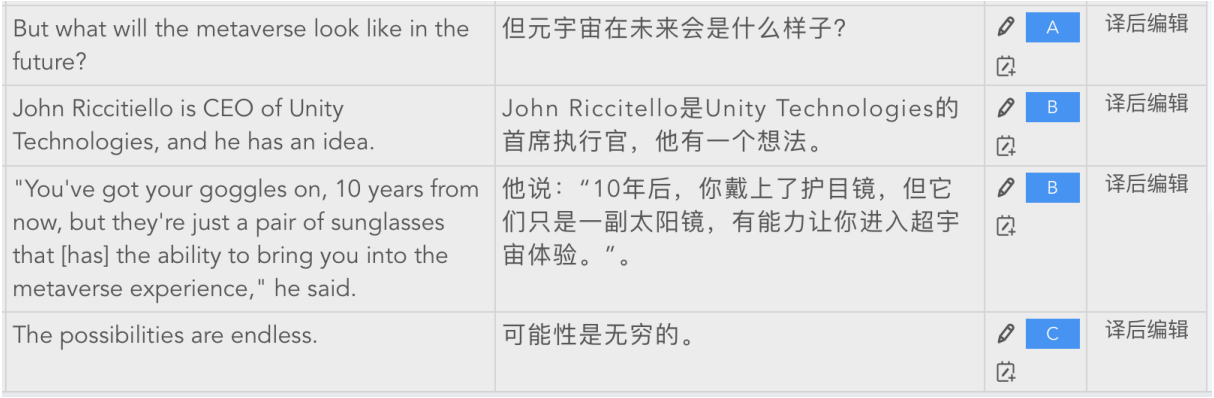


Figure 2: The YiCAT interface (Task 2, with QE information)

form. On the day of the experiment, a short guide on MTPE was first introduced to the participants, which covered key topics such as the concept of MTPE, differences between light and full MTPE, MTPE guidelines, MT quality assessment, and QE. Most importantly, participants were explicitly informed about the meaning of QE scores and encouraged to use them critically, since they may not always be accurate. This explanation was provided before both Task 1 and Task 2 to ensure a clear understanding of QE, even though QE information was only available in Task 2.

Participants were required to perform full MTPE according to GB/T 40036-2021: Translation services — Post-editing of machine translation output -Requirements[7], the Chinese national standard for post-editing. Then, a warm-up task was conducted by participants, followed by Task 1. A week later, similar procedures were followed for Task 2. Participants were again reminded of the MTPE guidelines, the interpretation of QE scores, and task requirements. Task 2 was conducted after a warm-up task. No time limits were imposed on the tasks, but participants were suggested to finish them as soon as possible. External resources, such as dictionaries, were prohibited.

Within two days of completing Task 2, twelve volunteers participated in one-on-one interviews. Participants were encouraged to share their experiences and perceptions of QE freely. The interviews followed a semi-structured outline, covering questions such as "when do you typically check QE scores (e.g., before reading the ST and MT; after reading the ST and MT; or after editing the MT)?"; "to what extent do you trust and rely on QE scores?"; and "do you think that adopting QE in post-editing tasks can increase efficiency?".

3.4 Data Processing and Statistical Analysis

The data analysis was conducted at the segment level using the statistical software R (R Core Team, 2024). Linear Mixed Effects Regression (LMER) models were employed to investigate the impact of QE on post-editing time. To address the first research question, a LMER model was built with task type (Task 1: without the aid of QE; Task 2: with the aid of QE) as the fixed effect. For the second research question, two additional LMER models were built. The first one included task type, MT quality, and their interaction as fixed effects. The second model included task type, translation expertise (students with CATTI Level 2 were classified as having higher expertise, while those with CATTI Level 3 were considered as having lower expertise), and their interaction as fixed effects. All models used post-editing time as dependent vari-

[7] https://www.gbstandards.org/China_standard_english.asp?code=GB/T%2040036-2021&id=49840

able, with participants and segments as random effects. Prior to model fitting, post-editing time was normalized by the number of words in the ST and transformed to approximate a normal distribution. Subsequently, the models were constructed, and their residuals were checked for normality and homoscedasticity.

It is important to note that, since there was only one segment being rated as low-quality by human raters, data pertaining to this segment was excluded from the model that included task type, MT quality, and their interaction as fixed effects. Therefore, the analysis of the interaction effect between MT quality and QE is limited to the cases of medium- and high-quality MT.

The interview data was transcribed and coded according to the outlined questions, serving as a complementary source to the post-editing experiment data in the current study. Due to the particular research focus and effort constraints, the analysis focused on participants' perceptions regarding the potential of QE to increase MTPE efficiency.

4 Results and Discussion

4.1 The Impact of QE on Post-editing Time

In order to assess the overall impact of QE, we first analysed the LMER model with task type as the main effect. As shown in Figure 3, Task 2 took less time than Task 1. To be more specific, the average time was 0.95s per word (SD=0.94) for Task 2, while it was 1.27s (SD=1) for Task 1. The main effect of the model was statistically significant (t=-2.34, p=0.02<0.05), suggesting that the use of QE information reduced post-editing time.

This reduction in post-editing time indicates the practical utility of QE information in enhancing translation efficiency. As mentioned previously, the significant impact of QE can be attributed to its potential to save translators time in evaluating MT quality and deciding whether to post-edit the outputs or discard them and translate from scratch. Additionally, QE may assist translators in quickly identifying and revising potentially erroneous segments, thereby prioritising and streamlining error correction. These findings align with previous research (Huang et al., 2014; Lee et al., 2021; Specia, 2011), which emphasised the role of QE in reducing processing time and enhancing workflow efficiency in post-editing tasks.

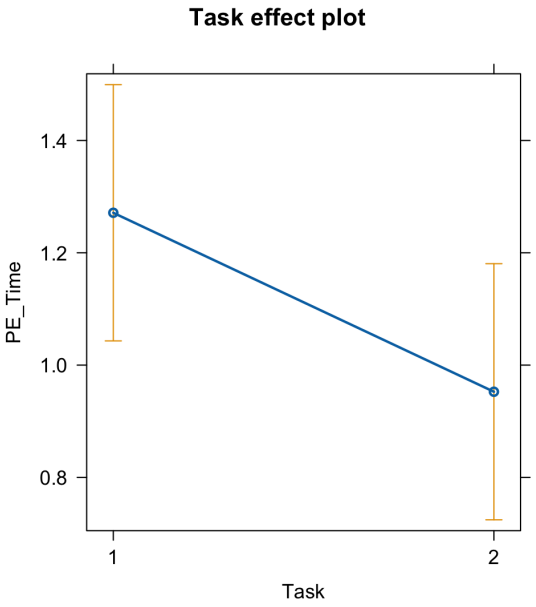


Figure 3: The effect of task type (Task 1: without QE, Task 2: with QE) on post-editing time

4.2 The Interaction Effect between QE and MT Quality

Having preliminarily established the significant impact of QE on post-editing time, we were interested in whether this effect remains consistent across different conditions. To address this question, we considered one important factor that could potentially influence post-editing time: the quality of MT outputs.

As illustrated in Figure 4 and supported by the model results, the interaction effect between MT quality and task type was not significant (t=0.62, p=0.54>0.05), indicating that the impact of task type on post-editing time remained consistent regardless of the MT quality levels. The effect of task type was significant (t= -2.13, p=0.04<0.05), with participants spending less time on Task 2 than on Task 1. MT quality also had a significant impact on post-editing time (t=-3.45, p=0.001<0.01). Specifically, MT outputs with lower quality led to longer post-editing time, which aligns with the findings of Gaspari et al. (2014), O'Brien (2011), and Tatsumi (2009).

The results suggest that post-editing with QE is consistently and significantly faster than without QE, no matter if the MT outputs are of medium or high quality. Our findings are partially consistent with those of Turchi et al. (2015), who observed that QE significantly increased post-editing speed when the HTER value was between 0.2 and 0.5.

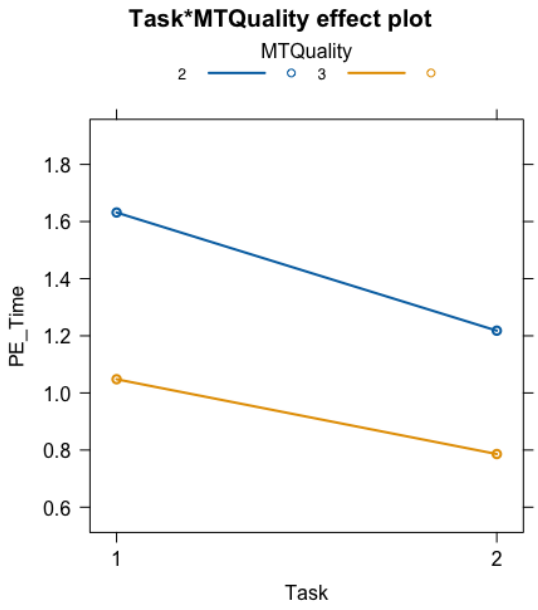


Figure 4: The interaction effect between MT quality (MTQuality=2: medium-quality MT, MTQuality=3: high-quality MT) and task type (Task 1: without QE, Task 2: with QE) on post-editing time

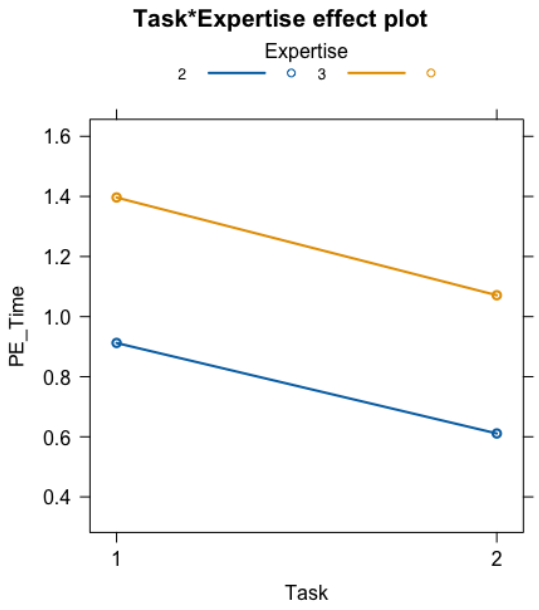


Figure 5: The interaction effect between translation expertise (Expertise=2: students with CATTI Level 2, Expertise=3: students with CATTI Level 3) and task type (Task 1: without QE, Task 2: with QE) on post-editing time

Although Turchi et al. (2015) did not categorise MT quality into high, medium, and low, they adopted a binary classification with a threshold of 0.4 to distinguish between editable and useless MT. In our study, both high- and medium-quality were considered "editable". Therefore, the 0.2 to 0.5 range identified by Turchi et al. (2015) overlaps to some extent with the quality levels examined in the current model.

These consistent efficiency gains suggest that QE can offer practical advantages across various scenarios. For instance, when dealing with high-quality MT outputs, presenting QE information may prevent translators from making unnecessary preferential edits. Such edits require certain effort and time but do not lead to increased translation quality and can sometimes even be detrimental (Koponen et al., 2019). Therefore, if translators know in advance that the segment they are working with is of high-quality, they are more likely to spend less time on it, thereby increasing post-editing efficiency. In the case of medium-quality MT outputs, QE can potentially help translators allocate their attention more effectively by identifying segments that are worthy of intervention. This allows them to concentrate on the task of editing itself, rather than second-guessing the overall quality of MT. Such a targeted approach can streamline the MTPE process, enabling translators to work more efficiently and effectively.

4.3 The Interaction Effect between QE and Translation Expertise

In addition to examining the quality of MT outputs, we also investigated the role of translation expertise in influencing the effectiveness of QE on post-editing time. The results indicated that the interaction effect between translation expertise and task type was not significant (t=-0.26, p=0.80>0.05). In other words, the impact of task type did not differ across student translators with varying levels of expertise. As shown in Figure 5, Task 2 required less time than Task 1 in both groups, suggesting that QE may have contributed to reduced post-editing time. The model results further indicated a marginally significant effect of task type (t=-1.97, p=0.05<0.1), pointing to a potential trend toward greater efficiency when QE information was available. In addition, translation expertise had a significant impact on post-editing time (t=3.46, p=0.001<0.01), with students with a higher level of expertise completing tasks more quickly than those with less expertise.

The results indicate that translation students, irrespective of their expertise levels, may have experienced similar improvements in speed from the

presence of QE information, although the observed advantages were only marginally significant. This finding aligns with Béchara et al.'s (2021) study, where nearly all professional translators across varying experience levels completed post-editing tasks more quickly with the aid of QE, except for one translator who maintained the same speed regardless of QE availability. One possible explanation is that QE provides explicit cues, so the cognitive processes involved in interpreting and utilising QE information may be straightforward, thus not necessitating advanced translation expertise. Moreover, participants in this study were informed that QE is not infallible and can make mistakes, and they were asked to engage with the information critically. It is therefore plausible that the students followed the instructions and integrated the QE information effectively, leading to productivity gains across the board. However, these findings warrant further validation, particularly through comparisons between professional and student translators, to confirm their generalisability.

4.4 Users' perceptions

This section focuses on participants' views on the potential of QE to increase MTPE efficiency. As summarised in Table 2, the interview data reveal a range of opinions, including some conflicting perspectives. Specifically, 66.7% (8) of the interviewees believed that QE could improve MTPE quality. Interestingly, while much of the previous literature has focused on QE's impact during the pre-processing stage (i.e. the process of evaluating MTPE's cost-effectiveness), participants in this study highlighted potential applications of QE in the later stages of MTPE. For example, interviewees reported using QE to check whether they had overlooked any MT errors, which is consistent with the results of Lee et al. (2021). Additionally, one participant used QE to validate her evaluation of MT quality, noting that this validation increased her confidence in the decisions regarding whether to edit MT outputs or not. However, among these eight interviewees, perceptions of QE's impact on MTPE speed were divided: half felt it helped them work faster, while the other half did not notice any meaningful improvement.

Notably, two participants perceived that QE had no impact on their MTPE processes, as they were very confident in their own assessment of MT quality. Finally, two students commented solely on QE's impact on speed without referencing its effect on quality. Their views were contradictory: one believed QE increased speed by highlighting potentially erroneous MT segments, while the other felt that QE slowed her down, citing distrust in its accuracy and a belief that the tool produced unreliable assessments. Although previous studies have not demonstrated that inaccurate QE negatively affects post-editing efficiency (Parra Escartín et al., 2017; Teixeira and O'Brien, 2017), particularly in comparison to working without QE, the interview data from this study suggest that poor QE accuracy may adversely impact users' experience and even reduce post-editing speed.

5 Conclusions and Future Work

Motivated by the potential benefits of QE in streamlining MTPE workflow, the current study preliminarily explored the usefulness of sentence-level QE in increasing post-editing speed and gathered student translators' views about its application. Three major findings emerged. First, QE significantly reduced post-editing time, and no significant interaction effects were found between QE and MT quality or between QE and translation expertise. In other words, the impact of QE remained consistent across MT outputs of medium and high quality and among students with varying levels of translation expertise. This stability implies that the advantages of QE in reducing post-editing time are likely to be broadly applicable. Second, the benefits of using QE in post-editing extend beyond highlighting problematic MT segments, it can also validate translators' own evaluations of MT quality and assist in quality checking. These findings shed new light on how translators can integrate QE information into the MTPE workflow to enhance overall efficiency. Finally, although this study did not explicitly examine the impact of QE's accuracy, interview data indicate a potential detrimental effect of inaccurate QE on post-editing processes. However, this finding requires further empirical validation.

This study has several limitations that open avenues for future research, particularly regarding the number and diversity of texts and participants, the range of factors considered, and the indicators used to measure MTPE efficiency. In future research, we aim to expand the sample size by including more participants and a wider variety of text types to better understand the conditions under which QE proves most beneficial. Comparisons between student and professional translators will also be

Views	N	Main Reasons
Increasing quality but not necessarily the speed	4(33.3%)	Validating translators' own evaluation of MT quality; assisting quality check
Increasing both quality and speed	4(33.3%)	Saving time and effort for more difficult translations; assisting quality check
No impact	2(16.7%)	A firm belief in translators' own evaluation of MT quality
Increasing speed	1(8.3%)	The ability of QE to highlight potentially erroneous MT
Decreasing speed	1(8.3%)	Low accuracy of QE; distrust of QE

Table 2: Users' perceptions of QE's potential in increasing MTPE efficiency

conducted to assess whether the benefits of QE differ when larger differences in expertise are present. Furthermore, low-quality MT outputs will be included to examine whether QE can still enhance post-editing efficiency in such cases. Other factors, such as the accuracy of QE and the score levels assigned by QE, will also be considered. Additionally, eye-tracking data, which can capture how translators allocate their cognitive resources when presented with QE information, will be collected to gain a more detailed understanding of QE's impact on the post-editing processes.

In conclusion, this study provides preliminary evidence for the usefulness of sentence-level QE in the MTPE context. Instead of simply saying "no" to QE, we should embrace its potential and investigate how to optimise its integration into MTPE workflows. As one interviewee aptly remarked, *"If we have access to such information, why not use QE?"* This perspective encapsulates the pragmatic value of QE and underscores the need for further exploration into its role in enhancing MTPE efficiency.

Acknowledgments

The work described in this paper was partially supported by the National Social Science Fund of China ("A Study on Quality Improvement of Neural Machine Translation", Grant reference: 22BYY042) and a grant from CBS Departmental Earnings Project of the Hong Kong Polytechnic University (Project title: Predicting Machine Translation Post-Editing Effort with Source Text Characteristics and Machine Translation Quality: An Eye-Tracking and Key-Logging Study; Project No.: P0051091).

References

Sergi Alvarez-Vidal and Antoni Oliver. 2023. Assessing mt with measures of pe effort. *Ampersand*, 11:100125.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

Hannah Béchara, Constantin Orăsan, Carla Parra Escartín, Marcos Zampieri, and William Lowe. 2021. The role of machine translation quality estimation in the post-editing workflow. *Informatics*, 8(3).

Scott A Crossley, Stephen Skalicky, and Mihai Dascalu. 2019. Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, 42(3-4):541–561.

Joke Daems, Sonia Vandepitte, Robert J. Hartsuiker, and Lieve Macken. 2017. Identifying the machine translation error types with the greatest impact on post-editing effort. *Frontiers in Psychology*, 8.

Guangrong Dai and Siqi Liu. 2024. Towards predicting post-editing effort with source text readability: An investigation for english-chinese machine translation. *The Journal of Specialised Translation*, 41:206–229.

Félix Do Carmo and Joss Moorkens. 2020. Differentiating editing, post-editing and revision. In *Translation revision and post-editing*, pages 35–49. Routledge.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221–233.

Federico Gaspari, Antonio Toral, Sudip Kumar Naskar, Declan Groves, and Andy Way. 2014. Perception vs. reality: measuring machine translation post-editing productivity. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas*, pages 60–72, Vancouver, Canada. Association for Machine Translation in the Americas.

Devin Robert Gilbert. 2022. *Directing Post-editors? Attention to Machine Translation Output That Needs Editing through an Enhanced User Interface: Viability and Automatic Application via a Word-Level Translation Accuracy Indicator*. Ph.D. thesis, Kent State University.

Fei Huang, Jian-Ming Xu, Abraham Ittycheriah, and Salim Roukos. 2014. Adaptive HTER estimation for document-specific MT post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–870, Baltimore, Maryland. Association for Computational Linguistics.

Yanfang Jia and Binghan Zheng. 2022. The interaction effect between source text complexity and machine translation quality on the task difficulty of nmt post-editing from english to chinese: A multi-method study. *Across Languages and Cultures*, 23(1):36–55.

Maarit Koponen, Leena Salmi, and Markku Nikulin. 2019. A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Machine Translation*, 33(1):61–90.

Hans P Krings. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*. Kent State University Press.

Dongjun Lee, Junhyeong Ahn, Heesoo Park, and Jaemin Jo. 2021. IntelliCAT: Intelligent machine translation post-editing with quality estimation and translation suggestion. 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 11–19, Online. Association for Computational Linguistics.

Joss Moorkens and Sharon O'Brien. 2013. User attitudes to the post-editing interface. In *Proceedings of the 2nd Workshop on Post-editing Technology and Practice*, Nice, France.

Joss Moorkens and Sharon O'Brien. 2017. *Assessing User Interface Needs of Post-Editors of Machine Translation*, pages 109–130. Routledge.

Joss Moorkens, Sharon O'brien, Igor AL Da Silva, Norma B de Lima Fonseca, and Fabio Alves. 2015. Correlations of perceived post-editing effort with measurements of actual effort. *Machine Translation*, 29:267–284.

Sharon O'Brien. 2011. Towards predicting post-editing productivity. *Machine translation*, 25:197–215.

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.

Carla Parra Escartín, Hanna Béchara, and Constantin Orasan. 2017. Questing for quality estimation a user study. *The Prague Bulletin of Mathematical Linguistics*, 108.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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.

Lucia Specia. 2011. Exploiting objective annotations for minimising translation post-editing effort. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, Leuven, Belgium. European Association for Machine Translation.

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, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine translation*, 24:39–50.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

Midori Tatsumi. 2009. Correlation between automatic evaluation metric scores, post-editing speed, and some other factors. In *Proceedings of Machine Translation Summit XII: Posters*, Ottawa, Canada.

Carlos Teixeira and Sharon O'Brien. 2017. The impact of MT quality estimation on post-editing effort. In *Proceedings of Machine Translation Summit XVI: Commercial MT Users and Translators Track*, pages 142–153, Nagoya Japan.

Marco Turchi, Matteo Negri, and Marcello Federico. 2015. MT quality estimation for computer-assisted translation: Does it really help? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 530–535, Beijing, China. Association for Computational Linguistics.

Lucas Nunes Vieira and Elisa Alonso. 2018. *The use of machine translation in human translation workflows: Practices, perceptions and knowledge exchange*. Institute of Translation and Interpreting.

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.

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.

Human- or machine-translated subtitles: Who can tell them apart?

Ekaterina Lapshinova-Koltunski, Sylvia Jaki, Maren Bolz, Merle Sauter
University of Hildesheim
Correspondence: lapshinovakoltun@uni-hildesheim.de, sylvia.jaki@uni-hildesheim.de

Abstract

This contribution investigates whether machine-translated subtitles can be easily distinguished from human-translated ones. For this, we run an experiment using two versions of German subtitles for an English television series: (1) produced manually by professional subtitlers, and (2) translated automatically with a Large Language Model (LLM), i.e., GPT4. Our participants were students of translation studies with varying experience in subtitling and the use of machine translation. We asked participants to guess if the subtitles for a selection of video clips had been translated manually or automatically. Apart from analysing whether machine-translated subtitles are distinguishable from human-translated ones, we also seek for indicators of the differences between human and machine translations. Our results show that although it is overall hard to differentiate between human and machine translations, there are some differences. Notably, the more experience the humans have with translation and subtitling, the more able they are to tell apart the two translation variants.

1 Introduction

Although Machine Translation (MT) has arrived in audiovisual translation somewhat later than in some other fields of translation, it has in fact come to play an important role in various translation forms such as subtitling, dubbing, etc. Idiomatic and enjoyable target texts are particularly crucial when it comes to the entertainment values that are typically associated with those types of translation, which is why there is skepticism among audiovisual translators concerning the quality of MT in this field (e.g. Jaki et al., 2024). On the other hand, the quality of MT has increased considerably over the last years, and it is common practice to use post-edited MT (MTPE), especially within the field of subtitling.

The question has therefore arisen whether MT subtitles are still recognisable as such. For this contribution, we analysed linguistic differences based on automatic annotation, as well as overlaps in words. This step involves a comparison of two translation variants using quantitative information on linguistic features. In addition, we asked human evaluators to recognise the method (manual or automatic) with which the subtitles at hand were produced, building on the results of Calvo-Ferrer (2023). For this step, students were asked to identify human and MT subtitles for an English TV series. Apart from the visible surface differences between the two translation variants measured by either linguistic information or human judgement, we are interested in further influencing factors, such as the quality of the subtitles or the test persons' level of expertise. For instance, it is interesting to know if dedicated instruction in subtitling increases the ability to recognise machine-translated subtitles and if other competences may play a role.

Thus, for our study, we formulate three research questions (RQs):

RQ1 Are there any differences between human and machine translation variants of the same subtitles?

RQ2 Does the quality play a role in the differentiation between human and machine-translated subtitles?

RQ3 Does the level of expertise play a role in the ability to tell apart human and machine translation?

In this study, we address the language pair English-German. Although both English and German are high-resource languages with much training data and existing MT solutions performing better than for other language pairs, we still believe

Proceedings of Machine Translation Summit XX Volume 1, pages 496–505, June 23-27, 2025

that looking into this language pair is important. The results of our study are particularly valuable for higher education institutions that train English-German subtitlers, since the information on the differences between MT and human subtitle translation is a great asset for this context.

The remainder of this paper is organised as follows: In Section 2, we give an overview of related work. Section 3 describes the data as well as the methodological design of this study. The results are presented in Section 4, which is organised along the RQs. We summarise the results as well as the limitations of this study, and we provide an outlook for future work in Section 5.

2 Related Work

2.1 MT technology for subtitling

Etchegoyhen et al. (2014)'s seminal work in the project SUMAT has marked a common strand of research in the automatic translation of subtitles that focuses on leveraging the quality of MT for subtitling, in part with feedback from professional subtitlers.

Over the time, neural machine translation (NMT) has taken the stand in the language industry as well as in research trying to boost these systems. Hiraoka and Yamada (2019), for example, obtained positive results for the translation pair Japanese-English by working with a set of pre-editing rules. Likewise, context has been increasingly considered in the improvement of MT systems. While Matusov et al. (2019) obtained positive results by including inter-sentence context, Vincent et al. (2024), in contrast, focused on including extra-textual information such as meta data into the MT model, working with MTCue (a multi-encoder transformer for contextual NMT). Their results imply that contextual data can improve the quality of MT for subtitles. Other researchers have chosen to use visual information to boost NMT performance, for example, Li et al. (2023) who successfully introduced SAFA, a new model for video-guided MT. As the focus of this study is not, however, a technological one, the remainder of the literature overview will go into more detail about the translation product, as well as the production and use of machine-translated subtitles.

2.2 Product-oriented studies

Hagström and Pedersen (2022) present a more product-oriented analysis of subtitles quality. They demonstrate a lower quality of subtitles since the 2020s, which they attribute to the increased use of MT. Other authors of product-oriented studies, in contrast, emphasise the general good quality of machine-translated subtitles, such as (Bellés-Calvera and Caro Quintana, 2021) for the English translation of the Spanish series *Cable Girls*. Martínez and Vela (2016) carry out an analysis of the quality in human- and machine-translated subtitles. They point out that although manual error analysis is time-consuming, it still provides interesting insights into the nature of human and machine translation in subtitling.

2.3 MT and subtitlers

Karakanta et al. (2022) focus on the subtitler's perspective and how MT influences their productivity. In this context, they test automatic subtitling (with MT as a part of automatic subtitling) with professional subtitlers and conclude that the subtitlers' post-editing experiences were "neutral to positive" (Karakanta et al., 2022, 9). Koponen et al. (2020) analyse the subtitling process in comparison between MT and HT and find that MTPE generally required fewer keystrokes that HT, but that there were considerable differences when it comes to language pairs, which emphasises the need for comprehensive research for a large variety of language pairs. Xie (2023)'s study of subtitler's effort in MTPE as part of automatic subtitling for the language pair English-Chinese concentrates particularly on the difference between videos with much information coming from the image in contrast to videos where most of the information stems from the verbal input. The author concludes that both require approximately the same time for MTPE, but that "the subtitlers spent more effort on revising spotting and segmentation than translation when they post-edited texts with more non-verbal information", and adds that MTPE was seen rather positively by the test persons (Xie, 2023, 63).

2.4 MT and end users

Other authors have focused on the end user's experience. For instance, Schierl (2023) shows in an analysis of Finnish and German subtitles that human translation in subtitles outperforms MTPE subtitles in terms of perceived quality, but that this does not mean that the end users need more time for reading MTPE subtitles (Schierl, 2023, 50). Calvo-Ferrer (2023) performs an experiment on the detectability of machine-translated subtitles for

the language pair English-Spanish. The approach is interesting as it combines a kind of Turing test with MT evaluation research. However, the experiment does not strictly address end users, as the test persons were 119 students of a translation study program. They were provided with eight clips with humorous content and were asked to classify those either as MT or HT. The results suggest that machine-translated subtitles have become difficult to identify. They also show that experience with translation seems to be a decisive factor: The fourth year students outperformed their fellow first year students in this classification task. The study also indicates that clips with poor subtitling quality are more frequently attributed to MT, and those of better quality to HT.

Our study directly builds on the results in Calvo-Ferrer (2023). We aim to find out whether we can find similar tendencies for the language pair English-German and if translation experience also plays a role. Whilst our experiment is designed to be comparable to the previous results by Calvo-Ferrer (2023) in the questions addressed, we also add linguistic analysis of the differences between human and machine translations, as well as the direct comparison of the outputs using the BLEU score (Papineni et al., 2002). Also, the data at hand differs from the data used in the previous research.

3 Methodology

3.1 Data

Subtitles For the experiment, we used freely available data provided on the homepage of IWLST (International Conference on Spoken Language Translation) for shared tasks on automatic subtitling (https://iwslt.org/2024/subtitling). IWLST obtained the data with the kind permission of ITV Studios, which has 60 labels in twelve countries and includes UK's largest commercial broadcaster (https://www.itvstudios.com/). The data set contains seven episodes of three different television series, with an approximate duration of seven hours in total, as well as their subtitles in English, German, and Spanish. To restrict the material, we selected seven clips that contained cultural references, puns, idioms, jargon-specific vocabulary, colloquial terms and elements of orality. In addition, we only chose one of the series and only scenes where the subtitles followed the subtitling guidelines provided by IWLST, therefore eliminating clips with subtitles up to three lines. Due to reasons of feasibility (as surveys need to be strictly limited in time, among other things to avoid fatigue effects), the material was again narrowed down to seven scenes, each with four to eleven subtitles in the German HT.

Automatic translation To produces machine-translated alternatives to the provided German subtitles, we used generative AI. More specifically, we performed tests with different models and on different web services: ChatGPT-4o mini on the Open AI web service[1], GPT-4o on a local university web service[2], as well as Meta LLaMA 3.1 8B Instruct[3] and GPT-o4 mini on ChatAI web service(Doosthosseini et al., 2024)[4]. In the end, we used the output of the best resulting translation as assessed by the authors of this paper, who have a background in linguistics, translation studies, and subtitling. Note that the outputs were not systematically compared with scores. Instead they were manually checked and the main attention was paid to the formal requirements for the subtitles as well as to linguistic accuracy. In our case, the best output was delivered with ChatGPT-4o mini. The results for this system were obtained by several prompts in German that we provide in Table 1, translated into English.

Prompt 2 was used to improve the result obtained from Prompt 1. For the human translation, we considered the subtitles provided by ITV as a gold standard, as we are dealing with human subtitles produced for a highly experienced broadcaster with a global outreach. None of the subtitles underwent any form of post-editing before the experiment, in order to avoid data manipulation. The subtitles were displayed to the test persons within the video clips, i.e., in their multimodal context.

Automatic annotation The collected human and machine translations were automatically annotated with parts-of-speech tags and syntactic functions with the help of the dependency parser using the Stanford NLP Python Library Stanza (v1.2.1)[5] with all the models pre-trained on the Universal Dependencies v2.5 datasets. We collected occurrence distribution of automatically tagged parts-of-speech (based on universal part-of-speech tags or UPOS) and selected syntactic functions that are assigned to

[1] https://chatgpt.com/
[2] Anonymised URL
[3] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[4] https://docs.hpc.gwdg.de/services/chat-ai/index.html
[5] https://stanfordnlp.github.io/stanza/index.html

System prompt: Your are a subtitler. For the translation of the English subtitles that I will be providing, please use the following rules for the output: The in- and out cues from the input as well as the time stamps are maintained like in a template; therefore they are not supposed to be changed. There is a maximum of 17 characters per second (including blanks). Each subtitle has a maximum of two lines (like in the input). There is a maximum of 42 characters per line (including blanks). The subtitles are produced for an American TV series from the Genre thriller or crime drama.
Prompt 1: Please translate the subtitles from English to German. Please stick to the rules indicated above.
Prompt 2: Please adjust the subtitles so that there are really only two lines per subtitle. Make sure they sound more colloquial and natural, like a conversation among colleagues. Don't forget to stick to the rules indicated above.

Table 1: Prompts used to translate subtitles into German.

the nominal category in the Universal Dependency classes (UD)[6], see de Marneffe et al. (2021) for more details. The occurrence of these categories was then compared between the two variants of translations.

3.2 Survey design

Building on the results by Calvo-Ferrer (2023), the survey was conducted among students at the University of Hildesheim in Germany. All the test persons were students of translation programs, i.e., they all had a very good command of English and native or near-native knowledge of German. One group was composed of 24 BA students. We assumed that they were not familiar with the art of subtitling yet. The other group consisted of 30 MA students that have already undergone instruction on subtitling, the hypothesis being that the students more experienced with subtitling may have less difficulty distinguishing between the machine-translated and the human subtitles. In order to control for the level of experience, both with MT and subtitling, we asked them how long they were studying, whether they had experience with subtitling and how often they were working with MT. We also asked students for their proficiency in English and German, mainly to understand potentially why grammatical mistakes or the like may have been overlooked in the subtitles. Please note that not all students finished the survey; fragmentary questionnaires were excluded from analysis. Consequently, the corpus of analysis consists of 21 answers from BA students and 25 from MA students.

The survey was implemented via Lime Survey[7], which allows for an exportation of data in excel format, for example. We also used the automatic shuffling function of Lime Survey to make sure that each participant would be confronted with seven video clips, with either machine-translated or human subtitles, respectively. For each of the clips, the participants had to indicate whether they thought they were dealing with human or machine-translated subtitles, how sure they were about their assessment in the respective cases, and how good they judged the quality of the subtitles to be. They were also provided with an open question for each of the clips where they had to indicate what their decision (MT vs. HT) was based on.

4 Results

4.1 RQ1: Differences between human and machine translation

We start with the first research question which concerns the differences between human and machine translations. To answer this question, we looked into both the frequency distributions of linguistic features in the two translation variants and into the human judgements collected in the survey.

Linguistic difference In terms of linguistic features, we analysed morpho-syntactic properties of the texts derived from the automatic analyses described in Section 3.1 above. We counted distributions of parts-of-speech (POS) and syntactic functions.

Figure 1 demonstrates the distributions of adjectives (ADJ), adpositions (ADP), adverbs (ADV), auxiliaries (AUX), connectives (CCONJ) and subjuncts (SCONJ), determiners (DET), common and proper nouns (NOUN, PROPN), pronouns (PRON),

[6] https://universaldependencies.org/u/dep/

[7] https://www.limesurvey.org

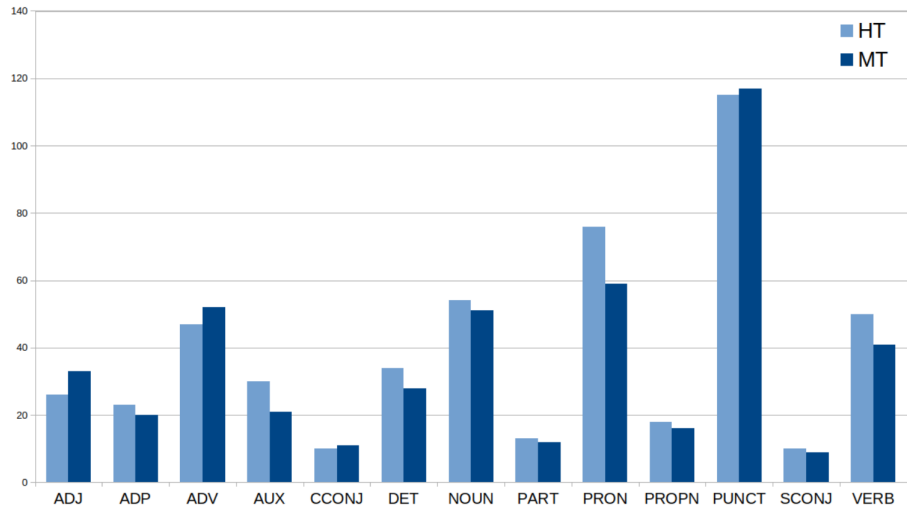


Figure 1: Distributions of parts-of-speech in human and machine translations.

verbs (VERB), particles (PART) and punctuation (PUNCT). The barplots reveal a number of differences in the distributions: While human translations contain more nouns, pronouns and verbs, machine-translated texts contain more adjectives and adverbs. However, the overall difference is not significant as confirmed by Pearson's chi-square test (p-value of 0.94).

We observe a similar tendency in terms of the distributions of the selected syntactic functions. They include nominal subjects (nsubj), direct objects (obj), indirect objects (obl), nominal modifiers (amod and nmod summarised as a-nmod in the figure), nominal modifiers functioning as appositions (appos), as well as adverbal modifiers (advmod), see Figure 2. It is obvious from the figure that the distributions of the categories are similar in both translation variants, with HT utilising more of those constructions. The most prominent difference is observed for the distribution of subjects, which prevail in human translations. However, the overall difference is not significant (p-value of 0.79).

Human judgements We proceed with the analysis of the survey results to see if students were able to recognise if the subtitles were translated manually or automatically. Table 2 represents the confusion matrix based on human judgements. The overall accuracy is relatively low (0.5). While human translations were recognised with 47.88% of precision, machine-translated texts seem to be slightly better identifiable - their recognition precision constitutes 51.59%. However, MTs have a lower true positive rate than human translations (0.49 vs. 0.51), which means that they were more frequently labeled as HTs.

true	HT	79	76
	MT	86	81
		HT	MT
		predicted	

Table 2: Confusion matrix: classification as HT and MT by test persons.

Table 3 illustrates the amount of correct judgements by BA and MA students. In general, the recognition rates were relatively low, and varied considerably between the different test items.

	BA		**MA**	
Clip	**total**	**in %**	**total**	**in %**
1	14	67	15	60
2	10	48	9	36
3	12	57	11	44
4	14	67	15	60
5	5	24	12	48
6	8	38	12	48
7	10	48	16	64

Table 3: Recognition rate per study degree.

For the BA students, MT was correctly recognised in 33 cases; HT was recognised correctly in 38 cases. MT was misinterpreted as HT 39 times, and HT as MT 37 times. The results differed considerably between the seven test items. For example, with two items, the subtitles were correctly classified as MT only once, respectively, while it was correctly classified as such 11 times with another test item. For the MA students, MT

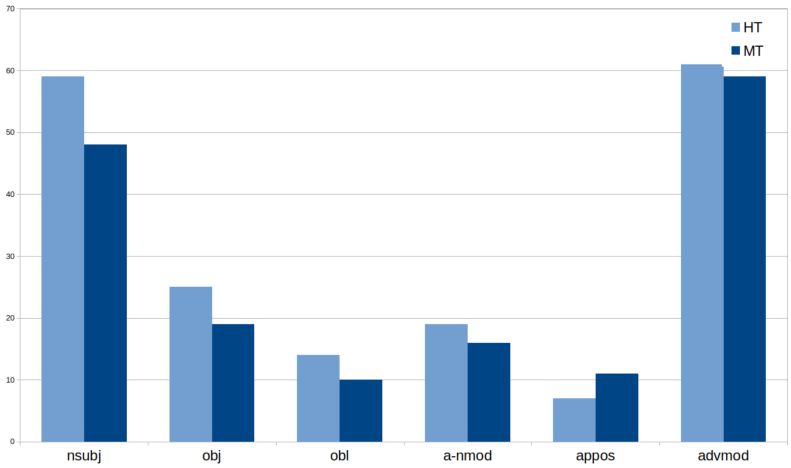


Figure 2: Distributions of syntactic functions in human and machine translations.

was recognised as such in 48 cases, HT in 41 cases. MT was erroneously identified as HT 47 times, and HT erroneously as MT 39 times.

4.2 RQ2: Role of quality

Next, we analysed if the quality of machine translation impacts the recognition rate. We also analysed if humans judge the quality of human and machine translations in a similar way.

MT quality As we were particularly interested in differences between human and machine translations and in their indicators, quality evaluation of MT is not the focus of this study. However, we calculated the automatic evaluation scores to get a general idea of their performance. Moreover, some scores, as e.g. BLEU score (Papineni et al., 2002), also provides information on the overlaps between human and machine translations. So, we used three metrics that can be calculated with the tools provided in MATEO (Vanroy et al., 2023), i.e., BLEU, ChrF (Popović, 2015), and TER (Snover et al., 2006). The numbers are reported in Table 5. All the scores point to dissimilarities between the two translation variants, as both BLEU and ChrF count the overlaps in ngrams between HT and MT (with ChrF taking into account also word order differences) and TER the edits needed fo MT to be overlapping with HT. This means that machine-translated texts in our data differ considerably from human translations in terms of word choices. In Table 4, we demonstrate an example from the data marking overlapping words in bold. As seen from this example, there is not much overlap in word choices between human and machine translations. At the same time, syntactic constructions, e.g. imperative in lines 6 and 7, seem to be similar, which coincides with our result on the linguistic differences measured with parts-of-speech and syntactic function distributions.

Using the data from the judgements by humans, we analysed if the BLEU score correlates with the misclassification cases, i.e., how many students labeled its machine-translated version as a human translation. As seen in Figure 3, we observe a negative correlation, which means that the quality of MT (at least the automatically evaluated quality) does not impact our test persons' decision and even the texts with lower scores can be identified as human translations.

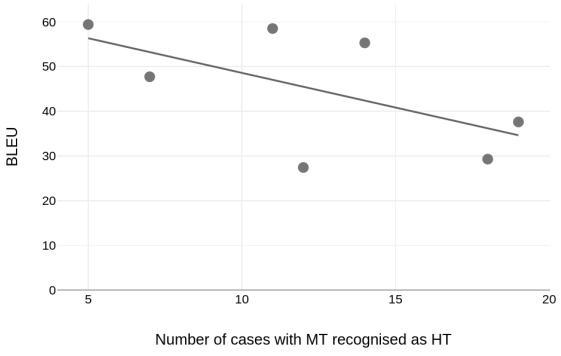


Figure 3: Correlation between human judgements and the quality of machine translation measured with BLEU.

Human quality judgements Participants were also asked to estimate the quality of the translation by labeling the test items with *very good*, *rather good*, *moderate*, *rather poor*, and *very poor*. An overview of the estimation per human and machine translations is given in Table 6. We indicate the percentage of answers normalised against the total

	human	machine
1	You had no idea.	
	Und ihr hattet keine Ahnung **?**	Hättest du nicht gedacht, oder **?**
2	Medal of Valor, Internal Affairs, cameras.	
	Tapferkeitsmedaille, interne Ermittlung, **Kameras**.	Medaille, Interne, **Kameras**... alles dabei.
3	It all seems a tad orchestrated, don't you think?	
	Scheint mir **ziemlich** viel Brimborium zu sein, findest du nicht?	Klingt alles irgendwie **ziemlich** inszeniert, oder?
4	If you think I had anything to do	
	Wenn **du** mir das anhängen willst,	Denkst **du**, ich hab' was damit zu tun,
5	with that, we can just step outside.	
	können **wir** gleich vor die Tür gehen.	dann klären **wir**'s draußen, okay?
6	Relax, Cut. This is not a John Wayne movie.	
	Krieg **dich** wieder ein, **Cut. Das ist kein** John-Wayne-Film.	Beruhig **dich, Cut. Das ist kein** Western.
7	Look at this. Everybody's doing the funny.	
	Sieh **dir das an.** Hier spielt **jeder** den Clown.	Schau **dir das an.** Jetzt macht **jeder** Witze.

Table 4: Example from the data: human (left) vs. machine (right) translation of Clip 1.

Clip	BLEU	ChrF	TER
1	15.7	27.4	84
2	9.7	29.3	75
3	13.6	37.6	75
4	27.6	47.7	62.5
5	32.9	55.3	60.6
6	31.5	58.5	53.7
7	26.8	59.4	65.5
Avg	**22.54**	**45.03**	**68.04**

Table 5: BLEU score per test item.

number of answers for MT and HT separately.

The test persons tended to rate the quality of the translations rather positively than negatively, but indicated a broad range of judgements: 44 times *very good* (13.84%), 135 *rather good* (42.45%), 97 times *moderate* (30.5%), 39 times *rather poor* (12.26%), and three times *very poor* (0.94%). Overall, both the machine-translated and manually produced outputs were rated similarly, with the only noticeable difference being a 6-per-cent higher rating for the human translations for the label *very good* (see Table 6). Nor do the results suggest that participants automatically associated those translations that they qualified as less good to be MTs, or those that they judged to be good to be HTs. This tendency can only be observed for Test Item 6, where the nine translations labeled as rather poor machine translation were all, in fact, human translations. Interestingly, out of the twelve HTs labeled with *very good*, eight were, in turn, MTs.

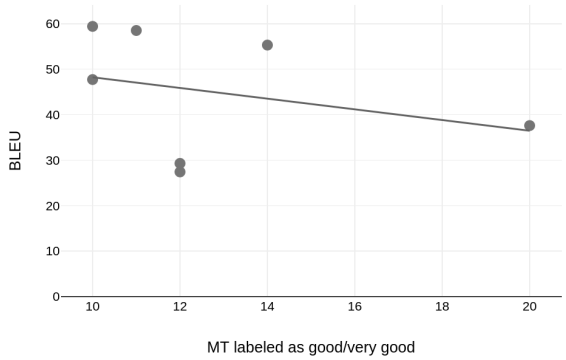


Figure 4: Correlation between human judgements and the quality of machine translation measured with BLEU.

We also analyse correlation between the BLEU score and the judgements by students. The latter is operationalised as the number of *good* and *very good* labels per MT version of the given video clip. As seen in Figure 4, the BLEU score [8] does not correlate with human judgements in our data, which again confirms the observation on RQ1 above.

4.3 RQ3: Role of the level of expertise

The level of expertise can be measured according to various criteria. All of the MA students in the experiment had had prior experience with the art

[8] We also tested correlation with ChrF and observed the same result as for BLEU.

MT vs. HT	Very good	Rather good	moderate	rather poor	very poor
MT	10.98	43.29	32.93	12.20	0.61
HT	16.56	40.76	28.03	13.38	1.27

Table 6: Ratings of the translations in per cent

of subtitling. At the same time, we included the experience with machine translation as a possible impacting factor too.

Figure 5 illustrates the number of correct judgements grouped by the study degree.

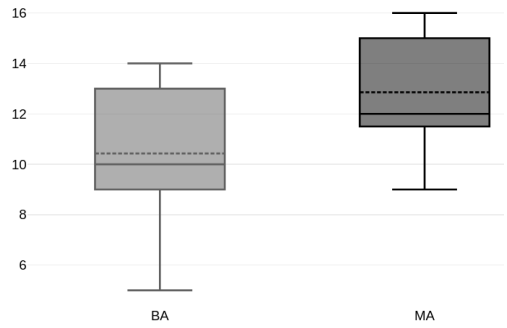


Figure 5: Correct judgements grouped by BA and MA study degree.

Overall, it was easier for the MA students to differentiate between human and machine translation, although the difference is not big.

To control for the **degree of familiarity with MT**, we asked students whether they worked with MT *very often*, *often*, *sometimes*, *rarely*, or *never*, with the hypothesis that it might be easier for students experienced with MT to distinguish between translations produced manually or automatically. The comparison between students who indicated that they worked with MT (1) *very often* or *often* and those who indicated (2) *sometimes* or *rarely* (*never* did not occur) did not produce any significant results: While there were 54 correct and 51 incorrect judgements for (1), it was 105 vs. 105 for (2). The amount of incorrect judgements was slightly higher when we singled out only those students who indicated that they rarely work with MT (with 27 correct and 29 incorrect answers). Therefore, it is fair to say that with the selection of students who are more experienced in MT, the amount of correctly identified translation variants is higher than the incorrect judgements. In contrast, the amount of correct judgements is lower than the amount of incorrect ones for the students who are not used to working with MT. However, the difference is only minor.

When it comes to the **level of confidence** in their answers, the difference between the correct and the incorrect answers is barely noticeable (possible answers: *very confident*, *rather confident*, *rather unsure*, *pretty unsure*): 19.02% of the correct and 14.47 % of the incorrect judgements were accompanied with *very confident*, 45.5% of the correct and 46.54% of the incorrect ones with *rather confident*, 30.06% of the correct and 32.08% of incorrect ones with *rather unsure*, and 5.52% of the correct and 6.96% the incorrect ones with *pretty unsure*.

5 Conclusion and Discussion

One aim of the present study was to see if machine-translated subtitles differ from human-translated ones. We used a number of analyses, including corpus-based frequency distribution of linguistic features, automatic quality scores, as well as human judgements. The overall results show that it is hard to differentiate between manually and automatically translated subtitles. Moreover, both translation variants seem to be similar in terms of the distribution of linguistic features such as parts-of-speech and syntactic functions. This points to structural similarities between the two outputs.

The main differences observed include word choice as indicated by the low BLEU score for machine translations, which implies that there are not so many n-gram overlaps between the two translation variants. Besides that, we showed that the BLEU score did not correlate with human judgements either, as texts with a lower BLEU score were more frequently labelled as human translations. Also, the calculated BLEU score does not necessarily reflect subtitle quality as it is perceived by humans, as our test persons classified the quality of machine translations as good and acceptable frequently, sometimes even more frequently than with the human-translated variants.

At the same time, it was interesting to see that the level of expertise measured by the advance in study program does play a role in the ability to correctly differentiate between human and machine translations of subtitles.

However, we are also aware of the limitations of this study. First of all, the number of the data that we included into the study (and also survey) is limited to seven texts (clips) only. This restriction was due to the requirements of the given settings: To avoid fatigue effects (which could have impacted the results), we decided beforehand that the survey time should be restricted to a maximum of 30 minutes. Given that watching the clips, making decisions and answering the questions takes a considerable amount of time, we could not collect data for more than the seven clips at hand.

We plan to extend the data to more clips. Although it is challenging to perform a survey with more texts, we would be able to perform a more extensive quantitative analysis of the linguistic differences between human and machine translations including automatic text classification.

Another drawback of this study is testing translation outputs with an LLM only. More machine-translated outputs, also those produced with traditional MT systems and with other LLMs than GPT is part of our future work. However, we are also aware of the problems of reproducibility, as the future results that build upon our findings may differ from those reported by us, as LLMs are regularly updated and are changing. Another problem of such systems is that we do not have any control over their training data. The dataset used for testing (the selected clips) is probably included into the training data of the LLMs at had, as the dataset is open source and freely available. Producing subtitle translation specifically for the survey would be a better scenario.

Besides that, this study does not provide a deep analysis of the subtitle quality. Although we mention some issues, we do not report on accuracy and other factors. Moreover, pragmatic factors such as transfer of emotions, sentiment, humour, etc. cannot be considered with the methodology applied. However, this can be analysed on the basis of our data in future work.

In future, we would also like to extend the test persons to more experienced groups and include professionals from the subtitling industry.

References

Lucía Bellés-Calvera and Rocío Caro Quintana. 2021. Audiovisual translation through NMT and subtitling in the netflix series 'cable girls'. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 142–148, Held Online. INCOMA Ltd.

José Ramón Calvo-Ferrer. 2023. Can you tell the difference? A study of human vs machine-translated subtitles. *Perspectives*, 32(6):1115–1132.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Ali Doosthosseini, Jonathan Decker, Hendrik Nolte, and Julian M. Kunkel. 2024. Chat AI: A Seamless Slurm-Native Solution for HPC-Based Services. *Preprint*, arXiv:2407.00110.

Thierry Etchegoyhen, Lindsay Bywood, Mark Fishel, Panayota Georgakopoulou, Jie Jiang, Gerard van Loenhout, Arantza del Pozo, Mirjam Sepesy Maučec, Anja Turner, and Martin Volk. 2014. Machine translation for subtitling: A large-scale evaluation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 46–53, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hanna Hagström and Jan Pedersen. 2022. Subtitles in the 2020s: The influence of machine translation. *Journal of Audiovisual Translation*, 5(1):207–225.

Yusuke Hiraoka and Masaru Yamada. 2019. Pre-editing plus neural machine translation for subtitling: Effective pre-editing rules for subtitling of TED talks. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 64–72, Dublin, Ireland. European Association for Machine Translation.

Sylvia Jaki, Maren Bolz, and Röther Sofie. 2024. KI-Technologien in der Audiovisuellen Translation. *trans-kom*, 17:320–342.

Alina Karakanta, Luisa Bentivogli, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2022. Post-editing in automatic subtitling: A subtitlers' perspective. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 261–270, Ghent, Belgium. European Association for Machine Translation.

Maarit Koponen, Umut Sulubacak, Kaisa Vitikainen, and Jörg Tiedemann. 2020. MT for subtitling: User evaluation of post-editing productivity. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 115–124, Lisboa, Portugal. European Association for Machine Translation.

Yihang Li, Shuichiro Shimizu, Chenhui Chu, Sadao Kurohashi, and Wei Li. 2023. Video-helpful multimodal machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4281–4299, Singapore. Association for Computational Linguistics.

José Manuel Martínez Martínez and Mihaela Vela. 2016. SubCo: A learner translation corpus of human and machine subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2246–2254, Portorož, Slovenia. European Language Resources Association (ELRA).

Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. Customizing neural machine translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy. 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.

Frederike Schierl. 2023. Reception of machine-translated and human-translated subtitles – a case study. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 42–53, Macau SAR, China. Asia-Pacific Association for Machine Translation.

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.

Bram Vanroy, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine translation evaluation online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500, Tampere, Finland. European Association for Machine Translation.

Sebastian Vincent, Charlotte Prescott, Chris Bayliss, Chris Oakley, and Carolina Scarton. 2024. A case study on contextual machine translation in a professional scenario of subtitling. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 561–572, Sheffield, UK. European Association for Machine Translation (EAMT).

Bina Xie. 2023. Machine translation implementation in automatic subtitling from a subtitlers' perspective. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 54–64, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Extending CREAMT: Leveraging Large Language Models for Literary Translation Post-Editing

Antonio Castaldo
University of Naples "L'Orientale"
University of Pisa
antonio.castaldo@phd.unipi.it

Sheila Castilho
Dublin City University
sheila.castilho@adaptcentre.ie

Joss Moorkens
Dublin City University
joss.moorkens@dcu.ie

Johanna Monti
University of Naples "L'Orientale"
jmonti@unior.it

Abstract

Post-editing machine translation (MT) for creative texts, such as literature, requires balancing efficiency with the preservation of creativity and style. While neural MT systems struggle with these challenges, large language models (LLMs) offer improved capabilities for context-aware and creative translation. This study evaluates the feasibility of post-editing literary translations generated by LLMs. Using a custom research tool, we collaborated with professional literary translators to analyze editing time, quality, and creativity. Our results indicate that post-editing LLM-generated translations significantly reduces editing time compared to human translation while maintaining a similar level of creativity. The minimal difference in creativity between PE and MT, combined with substantial productivity gains, suggests that LLMs may effectively support literary translators working with high-resource languages.

1 Introduction

Post-editing of MT has become an increasingly common service, given the cost-efficiency and good quality compromise that this practice offers. However, while several studies have confirmed that post-editing MT boosts productivity in terms of translation speed (Terribile, 2023), the benefits diminish significantly when dealing with poor-quality MT outputs (Guerberof Arenas, 2014; Sanchez-Torron and Koehn, 2016). This challenge is particularly pronounced for literary texts, where the final quality often suffers not only in terms of translation accuracy but also in the preservation of creativity, as discussed by Guerberof-Arenas and Toral (2020).

Recent LLM advancements have demonstrated significant improvements in handling context issues and figurative language to generate highly accurate and fluent translations. Unlike NMT systems that often tend towards generating translations that are either too literal or inaccurate, LLMs leverage large training data to generate context-aware translations less literally. Nevertheless, the extent to which they may support literary translators, without sacrificing creativity, remains underexplored.

In this study, we collaborated with four professional translators to evaluate the feasibility of post-editing literary translations generated by LLMs, focusing on three key aspects: editing time, translation quality, and creativity. We compare the performance of GPT-4, GPT-3.5, and a literary-adapted Mistral-7B model. We also developed a custom research tool called UniOr-PET (Castaldo et al., 2025) to collect detailed statistics on the editing process of a literary sci-fi novel.

Our findings reveal that post-editing LLM-generated translations between well-supported languages significantly reduces editing time compared to human translation while maintaining a similar level of creativity. As the difference in creativity scores between human and post-edited LLM translations appears to be minimal, our findings suggest that LLMs can serve as valuable tools for literary translators.

2 Related Work

Research on post-editing has traditionally centered on technical and commercial texts, where terminological consistency and turnaround time are often prioritized (Moorkens et al., 2018). However, trans-

Proceedings of Machine Translation Summit XX Volume 1, pages 506–515, June 23-27, 2025

lating creative works such as literature introduces unique challenges. NMT models have been shown to struggle with creative phraseological challenges, such as translating idiomatic expressions, where they often produce overly literal outputs.

Corpas Pastor and Noriega-Santiáñez (2024) highlighted these limitations, particularly in the context of literary texts. In contrast, Raunak et al. (2023) demonstrated that LLMs are capable of generating less literal and more contextually appropriate translations, especially when translating idiomatic expressions that tend to be generated with a higher level of abstraction, defined by the authors as "figurative compositionality". Further studies on idiomatic expression translation, particularly for the English-Italian language pair, have confirmed the high-quality results achieved by general-purpose LLMs (Castaldo and Monti, 2024). Their findings suggest that these models could address some of the shortcomings observed in NMT systems when translating literature, making them a promising tool for literary translation.

A study conducted by Guerberof-Arenas and Toral (2022) concluded that NMT was unable to handle the complex demands of translating literature or supporting literary translators effectively, resulting in low-quality outputs and diminished creativity. Their findings revealed the limitations of such models in preserving creativity during translation, becoming a constraint for the translator's creativity when used. Human translation (HT) consistently outperformed MT and PE in creativity, as evidenced by the annotation of units of creative potential. These findings align with the study by Castilho and Resende (2022), that showed how the features found in post-edited translations align more closely with the ones found in the MT output than in the HT. However, more recent advances in LLMs may shift this paradigm.

As demonstrated by Karpinska and Iyyer (2023) and Castilho et al. (2023), LLMs excel at leveraging training data to deal with context-related issues, which is critical for translating creative works that require discourse-level coherence and contextual understanding. Techniques such as in-context learning (Brown et al., 2020) and prompt engineering allow LLMs to maintain higher degrees of fluency, consistency, and stylistic fidelity compared to NMT systems. Finally, their ability to adapt to specific linguistic patterns and translation memories in real time, as shown by Moslem et al. (2023), further enhances their applicability in the creative translation domain, suggesting that LLMs could potentially overcome the creativity gap identified in NMT outputs, supporting professional translators in producing high-quality creative translations with context-aware terminology and accurate lexicon.

Drawing on Guerberof-Arenas and Toral (2022), in this study we consider creativity as a process that requires both originality and effectiveness (Runco and Jaeger, 2012). This implies that in order for a product to be creative, it needs not only to be novel but also of value, and therefore acceptable, for the context in which it is created. In Section 5, we will use the annotations of units of creative potential to reflect the original units introduced by the translators (novelty), and translation quality metrics as a proxy for the translation acceptability.

3 Methodology

We collaborated with four professional translators who specialize in literary and editorial translations to translate and post-edit excerpts from the novel "Oryx and Crake" by Canadian author Margaret Atwood (Atwood, 2004) from English into Italian. The novel was selected for its extensive use of playful and thought-provoking neologisms, vivid imagery, and richly detailed language, which present significant challenges in the translation process (Miller, 2019; Gurov, 2022; Noriega-Santiáñez and Corpas Pastor, 2023)

3.1 Participants

Each translator post-edited outputs of comparable length (roughly 2200 words), generated by three LLMs (see §3.2). We designed our study so that each translator contributes equally to the evaluation of the four models, rotating the chunks so that each translator works on three unique chunks, each generated by a different model. In this way, we minimize biases introduced by translator-specific behavior. We demonstrate our approach in Figure 1.

In addition, each translator produced a segment of the same excerpt translated from scratch. This experimental setup enabled us to collect fully post-edited translations for each model and a complete HT of the text for comparative analysis.

3.2 Models and Training

We employed three LLMs for generating the initial translations: `GPT-4`, `GPT-3.5`, and a literary-adapted `Mistral-7B` model, ordered by parameter

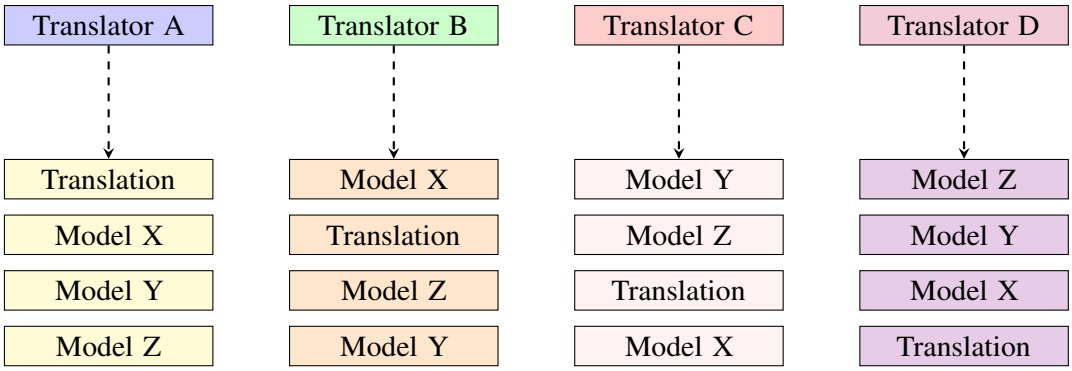


Figure 1: Each translator translates from scratch one chunk of original text (Translation) and post-edits a different chunk of each model's output (Model X, Y, Z), minimizing the translator's effect.

size. Access to the GPT models (OpenAI et al., 2024) was obtained through the OpenAI API,[1] as they both operate under closed-source licenses. In contrast, Mistral-7B (Jiang et al., 2023) was obtained as an open-source checkpoint, allowing us to fine-tune it locally for literary translation. Mistral-7B was fine-tuned on a curated corpus of modern literary works obtained from Opus Corpus (Tiedemann and Thottingal, 2020), for a total of 30,000 parallel segments. The model was fine-tuned for three epochs using Low-Rank Adaptation (Hu et al., 2021), a fine-tuning technique which injects small trainable matrices in the model's weights. The training corpus encompassed contemporary novels, short stories, and excerpts from science fiction and fantasy genres. The corpus was selected for its stylistic resemblance to the target text.

After fine-tuning, translation quality metrics and human inspection confirmed that Mistral-7B displayed improved handling of figurative language, idiomatic expressions, and higher accuracy. In terms of quality metrics, it achieved +4 points of corpus-level BLEU and +7 points of COMET as compared to its off-the-shelf counterpart.

3.3 Tools and Workflow

To facilitate the translation and post-editing process and collect meaningful data, we used two tools: our custom-built UniOr-PET and the established PET tool (Aziz et al., 2012).

UniOr-PET was designed specifically for this study, offering a browser-based platform that eliminates the need for software installation (see Figure 2). This feature addresses concerns often raised by translators regarding the inconvenience of downloading external applications, as is the case with

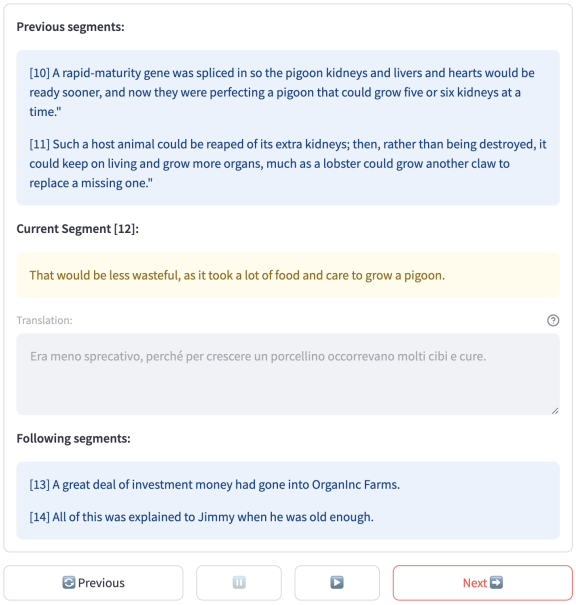


Figure 2: UniOr PET user interface

the PET tool. The tool records key metrics such as editing time, the number and types of edits, keeping track of insertions, and deletions. Similarly to the PET tool, UniOr PET gives the ability to read the texts, before recording editing time, making the results from both tools equally comparable. Translators could also save their work and revisit previously edited segments. The interface was configured to present the ST, LLM output, and an editable field, with a horizontal or vertical layout.

Recognizing the importance of context in literary translation (Nelson Jr., 1989; House, 2006), UniOr-PET also allowed translators to view a configurable number of preceding and following segments alongside the current one. This feature ensured that they could maintain consistency in tone, style, and narrative flow, an essential consideration when translating richly detailed texts, such as literature.

[1] https://openai.com

In addition to UniOr-PET, translators could opt to use the PET tool, which remains a popular choice for post-editing research due to its robust functionality and familiarity among professional translators, and researchers alike. Like its browser-based relative, PET captures data such as editing times and the types of edits made, providing a rich dataset for analysis. These tools provided translators with the flexibility to choose the interface that best suited their workflow preferences while allowing us to capture detailed post-editing data.

4 Results and Analysis

Thanks to the use of UniOr-PET and PET, we were able to collect significant data on each translation version providing foundation for a comparative analysis of the different models. More specifically, we have calculated quality metrics with BLEU (Papineni et al., 2002), ChrF (Popović, 2015) and COMET (Rei et al., 2020), which we average and normalize by time, as well as aggregated editing times. Finally, we compute Human-targeted Translation Edit Rate (Snover et al., 2006).

4.1 Editing Times

Source	Total
GPT-4	64.33
Mistral-60k	87.12
HT	115.68
GPT-3.5	119.74

Table 1: Editing Times (in Minutes)

Table 1 presents the aggregated total editing times (in minutes) for all translators and each part of the dataset. We find that editing time is shorter when post-editing outputs of the larger and best performing model used in our experiment, GPT-4. Interestingly, the literary-adapted Mistral model, despite its smaller size, demonstrated editing times significantly shorter than those for GPT-3.5. This suggests that domain adaptation, even in smaller models, can have a measurable impact on post-editing efficiency. These findings align with previous research indicating that better translation quality leads to reduced post-editing effort (Sanchez-Torron and Koehn, 2016; Zouhar et al., 2021).

The longest editing times were recorded when translating from scratch, which is expected since it requires significantly more technical (typing) effort than post-editing pre-generated MT outputs.

4.2 Human Translation Edit Rate (HTER)

Table 2 presents the HTER scores for the post-editing outputs from different MT systems. HTER is a widely used metric that quantifies the minimum number of edits required to improve an MT output when post-editing, where lower values indicate fewer required minimum edits. Therefore, HTER does not necessarily correspond to the actual number of edits, but rather represents an estimate of post-editing effort.

Source	T1	T2	T3	T4	Doc
GPT-3.5	44.4	41.9	62.2	31.8	**52**
GPT-4	50.4	66.5	52.2	29.9	54
Mistral-60k	66.1	66.0	71.5	54.5	**71**
HT	81.5	71.2	61.0	56.2	66
Total	242.4	245.6	247.0	172.4	226.85

Table 2: Human Translation Edit Rate. Lowest and highest HTER values are displayed in **bold**.

The results indicate varying levels of post-editing effort across the systems and across the four translators, with Translator 4 (T4) standing as an outlier when working with GPT models. This may be due to the adoption of a lighter form of post-editing, or an inclination to accept MT outputs considered sufficiently fluent and accurate.

We find that outputs from GPT-3.5 generally required the fewest edits, as reflected in the lowest HTER values among the systems. However, despite requiring fewer edits, post-editing outputs from GPT-3.5 took more time compared to the other models, as shown in Table 1. As both tools offer the possibility to read the texts, before performing translation, the results suggest that while the initial quality of GPT-3.5 translations was relatively higher, the type of edits required may have been more complex or time-consuming.

Interestingly, GPT-4 translations required more edits than GPT-3.5 but less overall editing time, indicating that its errors were likely easier to correct. Mistral-60k, while requiring more edits than GPT-3.5 and GPT-4, had comparable or shorter editing times, possibly due to simpler or more predictable error patterns. Translations from the ST show a significant difference from the reference translation, consistent with the lack of post-editing constraints.

As expected, we confirm a strong inverse correlation between HTER and quality metrics of the

original MT outputs, displayed in Figure 3, indicating that lower quality MT outputs require more post-editing efforts.

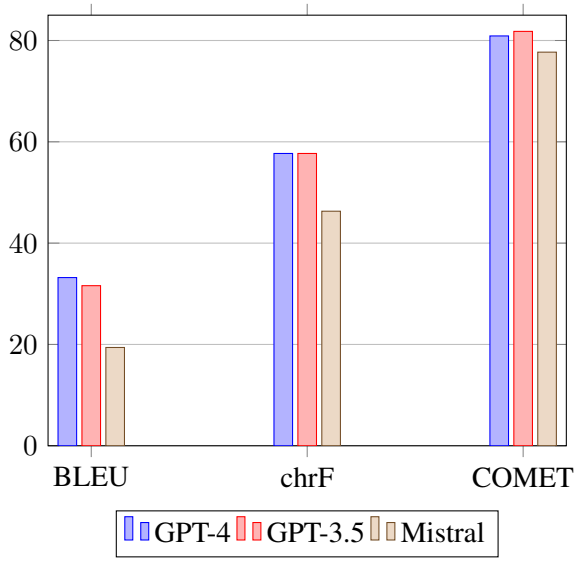


Figure 3: Quality metrics scores (BLEU, chrF, COMET) for different MT systems.

4.3 Quality-to-Time Ratio

Table 3 shows the normalized quality-to-time ratio for each MT system, calculated as the average of all quality metrics (BLEU, ChrF, and COMET) divided by the total editing time (Table 1). This ratio provides a measure of efficiency, combining the quality of the post-edited output with the time required to achieve it. Higher values indicate more efficient systems where higher-quality translations are achieved in less time.

Source	Ratio	BLEU	chrF	COMET
GPT-4	**0.38**	31.8	58.2	83.1
Mistral-60k	0.29	27.6	55.0	83.6
GPT-3.5	0.28	30.8	58.7	84.0
HT	0.23	27.1	54.4	80.5

Table 3: Quality-to-Time Ratio, calculated as the average of all quality metrics divided by the total editing time, along with BLEU, chrF, and COMET scores.

The results reveal that GPT-4 achieves the highest quality-to-time ratio (0.86), demonstrating the initial quality of the translation and the reduced post-editing effort, leading to good-quality post-edited translations in the shortest time.

Interestingly, Mistral-60k achieves the lowest ratio across the three models, despite requiring less editing time compared to GPT-3.5. This suggests that while Mistral translations may be quicker to edit, their initial quality presents challenges that limit their effectiveness in producing high-quality outputs efficiently, possibly resulting in the translator's decision to perform a lighter form of post-editing (Nitzke and Hansen-Schirra, 2021).

5 Creativity Annotation

To evaluate creativity in the post-edited translations and conduct a model-wise comparison, we annotated units of creative potential in the ST and creative shifts in the target texts (TT), that were originally generated by the three LLMs, and then post-edited by four translators.

Annotation Process. Our annotation framework follows the methodology proposed by Guerberof-Arenas and Toral (2022), where units of creative potential (UCPs) are defined as units that could invite creative deviations during post-editing, aimed at preserving or enhancing the creativity found in the ST, and creative shifts reflect the actual creative units introduced by translators during post-editing. Annotations were performed by two linguists with expertise in translation studies, who are native speakers of the target language and proficient in English. After annotating 10% of the dataset, inter-annotator agreement (IAA) was calculated to ensure the reliability of the annotations. The initial agreement, measured with Cohen's Kappa, was equal to $K = 0.35$ for Type Agreement and $K = 0.85$ for Span Agreement, due to disagreements primarily on the type of creative shift to assign, rather than the identification of the creative shifts themselves. Following a collaborative resolution process, we refined the annotation guidelines and calculated agreement on the final annotations, reaching a Type Agreement equal to $K = 0.57$ and a similarly high Span Agreement, equal to $K = 0.86$.

Creativity Score. A creative work must be both novel and acceptable, thereby achieving a balance between creativity and quality. In order to account for both novelty, as indicated by the number of creative shifts, and acceptability, as reflected by translation quality, we used `WMT22-COMET-DA` (Rei et al., 2022) for an automatic reference-based quality evaluation, and calculated the creativity score across the four translations.

In this study, we employ COMET as our primary metric for assessing translation quality, recogniz-

$$\text{Creativity Score} = \left(\frac{\#\text{CSs}}{\#\text{UCPs}} - \frac{\#\text{error points} - \#\text{Kudos}}{\#\text{words in ST}}\right) \times 100.$$

Figure 4: The original creativity score formula, that we started from to create our score.

ing that MQM would provide a more fine-grained evaluation of translation errors. Our decision to use COMET is motivated by its strong correlation with human judgments, as demonstrated in previous research (Rei et al., 2020; Kocmi et al., 2024), and by its practical advantage in automatic evaluation, in light of constraints related to time and resources. Having been trained on MQM-annotated datasets, COMET should effectively reflect the types of errors found in the outputs. Therefore, we integrate COMET in our creativity evaluation formula, as a proxy for translation acceptability.

Compared to the formula used in the original study, presented in Figure 4, we adapt the acceptability equation to accommodate the use of a quality metric, where higher means better, in place of the original error metric. Therefore, we multiply the creative shifts ratio by COMET scores, and then multiply by 100 to express it as a percentage. This allows us to reward creativity in proportion to quality, similarly to the original study. We present the new creativity score formula below.

$$\text{CS} = \left(\frac{\textit{Creative Shifts}}{\textit{UCPs}} \times \text{COMET}\right) \times 100 \quad (1)$$

5.1 Annotation Results

Table 4 summarizes the annotation results for each translation variant. For each system, we present the number of the creative shifts introduced by the translators, the COMET score, and the resulting creativity score, calculated with our new formula. A higher creativity score suggests a better balance between the introduced creative elements and the final translation quality.

System	CS Ratio	COMET	Creativity
HT	0.30	**0.85**	25.5%
GPT-3.5	0.24	0.84	20.1%
Mistral	0.30	0.83	24.9%
GPT-4	**0.32**	0.83	**26.5%**

Table 4: Creativity annotation results, where we display Creative Shifts ratio, COMET Score, and Creativity Score for each system.

6 Discussion

Taken together, our results show that a larger and more advanced model (GPT-4) generated translations that required fewer edits and resulted in a higher-quality post-edited translation, as resulted from the lower editing time and the higher quality-to-time ratio. The creativity score is also the highest, suggesting an interesting correlation between original MT quality and creativity in post-editing.

The domain-adapted Mistral-7B model also displayed promising performance, obtaining a quality-to-time ratio higher than the one obtained by the larger GPT-3.5, requiring more edits but a significantly lower editing time, while obtaining a similar creativity score. In this case, we find that Mistral's creativity comes at the cost of increased post-editing effort. HT, despite requiring a significantly higher editing time, is the most accurate translation variant according to COMET scores and it presents a high creativity score that is very similar to the post-edited texts.

In Table 5 we present two segments for each translation version with the highest and lowest post-editing effort, as measured by HTER. In displaying the segments, we ignore cases where the HTER is equal to zero due to translators not making any changes to the MT output. The examples reveal several interesting patterns. In some cases, the translators decided to merge or split certain sentences. Extensive edits were made in segments containing UCPs, as in the second example for GPT-3.5. Similarly, we find several edits where the original MT quality was particularly low, as seen in the second segment from the Mistral model. Interestingly, we find that where the MT systems failed to render neologisms effectively, translators were forced to produce a creative alternative, effectively improving the creativity of the translation.

Overall, we find that the creativity score does not differ significantly between the four models, as both the number of identified creative shifts and the quality metrics are similar across all translation variants. These findings are in contrast with what was found in the original study, where the difference between the two modalities (HT and PE) was substantial and HT was found to be notably more

creative than their post-edited variant. We speculate that the higher and more fluent MT quality given by LLMs may be of less constraint to the translator in the post-editing process, leading to equally creative translations.

7 Conclusion

In this study, we investigated the potential of LLM-based post-editing in the literary domain, comparing a literary-adapted Mistral model with GPT-4 and GPT-3.5. By collaborating with four professional literary translators, we collected detailed data on editing times, error rates, and post-editing efficiency, using our custom-built tool UniOr-PET. We demonstrate the contributions that LLMs can make in literary post-editing workflows, bridging the gap between productivity and creativity.

Our findings highlight two important benefits granted by the adoption of LLMs. First, we demonstrate that, in the context of our study, creativity does not present a significant difference between human translation and post-edited LLM translations. The marginal difference in creativity between the four translation variants suggests that the post-edited outputs may preserve creativity effectively. This may be due to the more fluent and higher-quality outputs given by the original MT versions, that represent less of a constraint to the translators, compared to NMT outputs.

Second, we observe a clear productivity gain in post-editing compared to human translation, even when post-editing translations generated by a smaller model. Given that the creativity gap is relatively small across translation variants, the productivity gains may offset the minor differences in creativity, achieving similarly creative translations with significantly less effort and time.

Finally, we reinforce the potential of fine-tuning techniques for literary MT workflows, demonstrating that even by adopting a small literary-adapted model, it is possible to achieve a good balance between translation quality and efficiency.

8 Limitations

One of the main limitations of this study is that our data collection process involved only four translators working in a single relatively well-resourced language pair and a relatively short literary excerpt. Further studies, on a larger scale, are required to investigate the possible correlations between creativity and other metrics. It is also worth mentioning that although our study follows established proxies for measuring creativity, these should be verified with a reception study, as suggested by Guerberof-Arenas and Toral (2020).

For the acceptability score, meant to balance creativity by translation quality in the post-edited texts, we used COMET scores in place of human evaluation. While COMET has shown strong correlations with human judgment, it remains an automated metric and may not fully capture the extent of literary translation quality.

Finally, while our literary-adapted Mistral model showed promising performance, its fine-tuning was performed using a modest-sized corpus, leaving open the way for further experimentation.

8.1 CO2 Emission Related to Experiments

Experiments were conducted using Amazon Web Services in region eu-west-1, which has a carbon efficiency of 0.62 $kgCO_2eq/kWh$. A cumulative of 3 hours of computation was performed on hardware of type RTX A6000 (TDP of 300W).

Total emissions are estimated to be 0.56 $kgCO_2eq$ of which 100 percents were directly offset by the cloud provider.

Estimations were conducted using the Machine Learning Impact calculator presented in Lacoste et al. (2019).

Acknowledgments

We thank the two annotators who took part in this study. Part of this work has been funded by the Italian National PhD programme in Artificial Intelligence, partnered by University of Pisa and University of Naples “L’Orientale", through a doctoral grant (ID 39-411-24-DOT23A27WJ-6603) established by Ex DM 318, of type 4.1, co-financed by the National Recovery and Resilience Plan. The second and third authors 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

Margaret Atwood. 2004. *Oryx and Crake*. Number v.1 in The MaddAddam Trilogy Ser. Knopf Doubleday Publishing Group, New York.

Wilker Aziz, Sheila Castilho, and Lucia Specia. 2012. PET: a Tool for Post-editing and Assessing Machine

Translation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3982–3987, Istanbul, Turkey. European Language Resources Association (ELRA).

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, pages 1877–1901, Red Hook, NY, USA. Curran Associates Inc.

Antonio Castaldo, Sheila Castilho, Joss Moorkens, and Johanna Monti. 2025. Unior PET: An Online Platform for Translation Post-Editing. In *20th Machine Translation Summit: Products and Projects track*, Geneva, Switzerland. European Association for Machine Translation.

Antonio Castaldo and Johanna Monti. 2024. Prompting Large Language Models for Idiomatic Translation. In *Proceedings of the First Workshop on Creative-text Translation and Technology*, pages 37–44, Sheffield, UK. Accepted: 2024-06-19T21:00:05Z.

Sheila Castilho, Clodagh Quinn Mallon, Rahel Meister, and Shengya Yue. 2023. Do online Machine Translation Systems Care for Context? What About a GPT Model? In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 393–417, Tampere, Finland. European Association for Machine Translation.

Sheila Castilho and Natália Resende. 2022. Post-Editese in Literary Translations. *Information*, 13(2):66. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Gloria Corpas Pastor and Laura Noriega-Santiáñez. 2024. Human versus Neural Machine Translation Creativity: A Study on Manipulated MWEs in Literature. *Information*, 15(9):530.

Ana Guerberof Arenas. 2014. Correlations between productivity and quality when post-editing in a professional context. *Machine Translation*, 28(3):165–186.

Ana Guerberof-Arenas and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, 9(2):255–282.

Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*, 11(2):184–212.

Andrey Gurov. 2022. Literary Translation as An Insurmountable Obstacle for Neural Networks. *SSRN Electronic Journal*.

Juliane House. 2006. Text and context in translation. *Journal of Pragmatics*, 38(3):338–358.

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 [cs].

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. *arXiv preprint*. Issue: arXiv:2310.06825 arXiv:2310.06825 [cs].

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. Issue: arXiv:2304.03245 arXiv:2304.03245 [cs].

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the Metrics Maze: Reconciling Score Magnitudes and Accuracies. Issue: arXiv:2401.06760 arXiv:2401.06760 [cs].

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Tristan Miller. 2019. The Punster's Amanuensis: The Proper Place of Humans and Machines in the Translation of Wordplay. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 57–65, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Joss Moorkens, Antonio Toral, Sheila Castilho, and Andy Way. 2018. Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7(2):240–262. Publisher: John Benjamins Publishing Company.

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.

Lowry Nelson Jr. 1989. Literary Translation. *Translation Review*, 29(1):17–30. Publisher: Routledge.

Jean Nitzke and Silvia Hansen-Schirra. 2021. *A short guide to post-editing (Volume 16)*. Language Science Press.

Laura Noriega-Santiáñez and Gloria Corpas Pastor. 2023. Machine vs Human Translation of Formal Neologisms in Literature: Exploring E-tools and Creativity in Students. *Tradumàtica tecnologies de la traducció*, (21):233–264.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, and Greg Brockman. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

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.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan Awadalla. 2023. Do GPTs Produce Less Literal Translations? ArXiv: 2305.16806.

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.

Mark A. Runco and Garrett J. Jaeger. 2012. The Standard Definition of Creativity. *Creativity Research Journal*, 24(1):92–96.

Marina Sanchez-Torron and Philipp Koehn. 2016. Machine Translation Quality and Post-Editor Productivity. In *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*, pages 16–26, Austin, TX, USA. The Association for Machine Translation in the Americas.

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.

Silvia Terribile. 2023. Is post-editing really faster than human translation? *Translation Spaces*, 13(2):171–199. Publisher: John Benjamins Publishing Company.

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.

Vilém Zouhar, Martin Popel, Ondřej Bojar, and Aleš Tamchyna. 2021. Neural Machine Translation Quality and Post-Editing Performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10204–10214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Model	Type	Text
GPT-3.5 *(Lowest HTER)*	ST	*But he hadn't wet his bed for a long time...*
	HT	*Eppure era un pezzo che non bagnava il letto...* yet was a while that not wet the bed...
	MT	*Ma non aveva bagnato il letto da molto tempo...* but not had wet the bed since much time...
	PE	*Eppure era da un pezzo che non bagnava il letto...* yet was since a while that not wet. the bed...
GPT-3.5 *(Highest HTER)*	ST	*Some cheap do-it-yourself enlightenment handbook*
	HT	*Uno scadente manuale di auto rivelazione* a poor manual of self revelation
	MT	*Una specie di manuale economico per l'illuminazione* a kind of manual cheap for the.enlightenment
	PE	*Una specie di manuale a poco prezzo per raggiungere l'illuminazione* a kind of manual at little price to reach the.enlightenment
GPT-4o *(Lowest HTER)*	ST	*All of this was explained to Jimmy when he was old enough.*
	HT	*Tutto questo fu spiegato a Jimmy quando fu abbastanza grande.* all this was explained to Jimmy when was sufficiently big.
	MT	*Tutto questo fu spiegato a Jimmy quando era abbastanza grande.* all this was explained to Jimmy when he.was sufficiently big.
	PE	*Tutto questo venne spiegato a Jimmy quando fu abbastanza grande.* all this came explained to Jimmy when was sufficiently big.
GPT-4o *(Highest HTER)*	ST	*She's got her own ideas.*
	HT	*Ha le sue idee.* has the her ideas.
	MT	*He le sue proprie idee.* he the his own ideas.
	PE	*Abbiamo opinioni diverse sulla cosa.* we.have opinions different on.the thing.
Mistral *(Lowest HTER)*	ST	*Ramona was one of his dad's lab technicians.*
	HT	*Ramona era uno dei tecnici di laboratorio di suo padre.* Ramona was one of.the technicians of laboratory of his father.
	MT	*Ramona era una delle tecniche del laboratorio del padre.* Ramona was one of.the technicians.FEM of.the laboratory of.the father.
	PE	*Ramona era una dei tecnici del laboratorio di suo padre.* Ramona was one.FEM of.the technicians of.the laboratory of his father.
Mistral *(Highest HTER)*	ST	*They called the cities the pleeblands.*
	HT	*Chiamavano le città plebopoli.* they.called the cities plebopolis.
	MT	*Chiamavano le città le plebe.* they.called the cities the plebs.
	PE	*Si riferivano alle città chiamandole terre di plebelandia.* they referred to.the cities calling.them lands of plebelandia.

Table 5: Examples of source text (ST), human translation (HT), machine translation (MT), and post-edited output (PE) for GPT-4o, GPT-3.5, and Mistral, showing segments with glosses and the lowest and highest post-editing effort as measured by HTER.

To MT or not to MT: An eye-tracking study on the reception by Dutch readers of different translation and creativity levels

Kyo Gerrits and Ana Guerberof Arenas
Center for Language and Cognition (CLCG), University of Groningen
k.gerrits@rug.nl and a.guerberof.arenas@rug.nl

Abstract

This article presents the results of a pilot study involving the reception of a fictional short story translated from English into Dutch under four conditions: machine translation (MT), post-editing (PE), human translation (HT) and original source text (ST). The aim is to understand how creativity and errors in different translation modalities affect readers, specifically regarding cognitive load. Eight participants filled in a questionnaire, read a story using an eye-tracker, and conducted a retrospective think-aloud (RTA) interview. The results show that units of creative potential (UCP) increase cognitive load and that this effect is highest for HT and lowest for MT; no effect of error was observed. Triangulating the data with RTAs leads us to hypothesize that the higher cognitive load in UCPs is linked to increases in reader enjoyment and immersion. The effect of translation creativity on cognitive load in different translation modalities at word-level is novel and opens up new avenues for further research. All the code and data are available at https://github.com/INCREC/Pilot_to_MT_or_not_to_MT.

1 Introduction

Recently, publishing houses have been more vocal about the use of machine translation (MT) in their translation process (Klemin, 2024), arguing that the output quality is good enough to post-edit certain genres considered less literary such as crime or romance novels in certain language combinations. However, no data has been provided to illustrate not only the impact on translators' livelihood and sustainability of high-quality literary translations but also what the impact on the readers of these books might be. Our research focuses on the latter, i.e. we seek to explore the effects of the use of MT-mediated texts in literary translation.

Having this goal in mind, this study uses materials and research methods from an existing study by Guerberof-Arenas and Toral (2024) that explored how different reading modalities (MT, PE, HT and ST) affect Dutch readers regarding engagement, enjoyment, and translation reception. We include two new methodological parts: a) an eye-tracking device to obtain granular data on readers' attention (cognitive load) and b) retrospective think-aloud (RTA) interviews to understand the differences in the readers' experiences when reading these texts. We focus on the effect of creativity across the different modalities by analysing the reception of units of creative potential. A unit of creative potential is a word or group of words that present a problem to the translator that requires a higher level of creativity, see Section 2.2

With this experiment, our main aim is to find a methodological framework to measure cognition and creativity, which, to our knowledge, has not been attempted before. To test this methodology, we guide the experiment with the following research questions: RQ1: Do readers have a higher cognitive load in units of creative potential than in other regular parts of the sentence? RQ2: Do readers process these units differently according to the translation modality? RQ3: Do readers process the translators' solutions differently according to the level of creativity? RQ4: Do errors in a segment increase the cognitive load of the reader?

2 Reading and translation reception

2.1 Eye-tracking and reading studies

Eye-tracking is a common method for measuring cognitive load. Even before technology was used for measuring the eye-movements, researchers already thought cognitive effort influenced eye movements during reading—-now known as the cognitive-control hypothesis (Rayner and Reingold, 2015). Yet there were also doubts about

Proceedings of Machine Translation Summit XX Volume 1, pages 516–537, June 23-27, 2025

eye-tracking methods. As average fixations are around 250 ms, some thought that it would be too little time for lexical processing (Chanceaux et al., 2012). Others also believed eye movements were largely caused by involuntary oculomotor movements (Yarbus, 1967). However, numerous studies have since shown that eye movement is in fact heavily influenced by cognitive effort (Reichle and Reingold, 2013; Schotter et al., 2017; Madi et al., 2020; Dias et al., 2021).

Many reading studies focus on comprehension; although intuitively we might think that lower reading times correlate with better comprehension, research shows that increased fixation duration and count might indicate higher comprehension levels (Mézière et al., 2023; Southwell et al., 2020; Wonnacott et al., 2016). Furthermore, studies looking at reading effort in literary texts find the more literary and style-heavy a text is considered to be, the more and longer the fixations are (Corcoran et al., 2023; Fechino et al., 2020), especially foregrounded elements (Jacobs, 2015; Müller et al., 2017).[1] Torres et al. (2021) also found that this increase in cognitive load also seems related to increased immersion in and engagement with the text, as reported by participants.

2.2 Creativity in translation

A clear conceptualisation of creativity in translation studies is introduced by Kussmaul in his seminal work (1991; 1995; 2000a; 2000b). This is further operationalised by Bayer-Hohenwarter (2009; 2010; 2011; 2013). They describe a creative translation as "involv[ing] changes (. . .) when compared to the source text, thereby bringing something that is new and also appropriate" (Bayer-Hohenwarter and Kussmaul, 2020, p. 312). Guerberof-Arenas and Toral (2020) further develops the concepts of unit of creative potential (UCP) and creative shift (CS) to measure creativity in literary translation and see the impact creativity has on readers. UCPs are problematic units in the source text that translators cannot translate routinely and for which they have to use problem-solving abilities, that is, their creative skills (Bayer-Hohenwarter, 2011). CSs are translated UCPs in which the translation deviates from the original, in contrast to Reproductions, where UCPs do not deviate from the original in structure or where there is already a coined translation (Guerberof-Arenas and Toral, 2020). This is illustrated in Figure 1. Our study uses this conceptualisation of creativity to annotate the translations, which allows us to analyse cognitive load (eye-tracking) across creativity (RQs 1, 2 & 3).

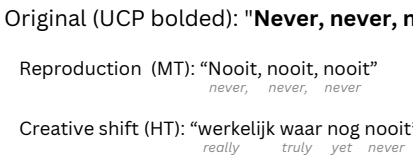


Figure 1: An example of UCP, Reproduction, and CS from the experiment, including word-level glosses.

2.3 Reading in the Netherlands

Dutch readership has some peculiarities worth mentioning regarding the cohabitation of English and Dutch languages. Recent market research shows that sales of foreign language books have increased 124% since 2020, accounting for 25% of all sales in 2024 (KVB Boekwerk, 2025), with the majority of these being English. Many readers read or even prefer reading books in the original language: 41% of Dutch readers regularly read books in another language, of which 77% are books in English; 29% even prefer English-language books to Dutch ones (KVB Boekwerk, 2022). This is interesting both for the potential influence of English phrases or structures on readers, and because English might be their default when opting to buy a book.

2.4 Literary translation and MT

A known issue in literary MT is the presence of errors, despite continuous improvements (Stasimioti et al., 2020, Matusov, 2019). Recent developments in neural machine translation (NMT) and large language models (LLMs) show to have increased MT quality (Son and Kim, 2023), but studies looking at LLMs for literary translation still find numerous errors in the MT output (Zhang et al., 2024), even when prompted to correct this MT output (Egdom et al., 2024; Macken, 2024). In the particular case of English-Dutch MT: Fonteyne et al. (2020) and Tezcan et al. (2019) look at an English-to-Dutch NMT version of Agatha Christie's *The Mysterious Affair at Styles* and find that 44% sentences had no errors while 56% still contain errors. A follow-up study by Webster et al. (2020) looks at four different novels and finds a much higher number of incorrect sentences in the NMT output, 77% with errors

[1] Stylistic devices that emphasize certain parts of the text to increase the impact on the reader.

vs 23% without. They argue that this difference could be related to the ST linguistic complexity.

If PE is considered, some studies indeed show a decrease in errors when compared to raw MT output (Guerberof-Arenas and Toral, 2020; 2022). PE might also provide a lower cost for industry or shorten the translation times (Toral et al., 2018), although this also depends on the desired final quality. However, PE is not without challenges. One of these is that the MT output tends to prime post-editors and this results in a final text that is syntactically, semantically and stylistically closer to MT than to HT in both technical and literary texts (Toral, 2019; Daems et al., 2024; Macken et al., 2022; Kolb, 2021; Castilho and Resende, 2022). PE has also been shown to reduce literary style and authorial voice compared to HT (Kenny and Winters, 2020; Mohar et al., 2020; Şahin and Gürses, 2019). Last but not least, translators have expressed their dislike of using PE, preferring to translate from scratch for creative purposes (Moorkens et al., 2018; Daems and Macken, 2019).

2.5 Translation reception and MT

Translation reception concerns how readers react to a translation, such as emotional (e.g. enjoyment) and cognitive responses (confusion, reading times). There are relatively few studies on translation reception (Walker, 2021). Some studies focus on the effect of errors in non-literary MT: Kasperavičienė et al. (2020) and Stymne et al. (2012), for instance, use eye-tracking to analyse the effect of errors in newspaper articles and political discussions respectively. They find that total fixation duration and fixation count are higher on sections that contain errors. Whyatt et al. (2024), also using an eye-tracker, analyse the reception of newspaper translations of low and high quality and find that participants spend more time on sentences of lower quality and sentences with errors than in those without. These studies suggest that a text with more errors and of lower quality garners more cognitive load, at least in non-literary texts.

Others look at literary MT. Colman et al. (2022), for instance, explore the reception of an English-into-Dutch MT version of Agatha Christie's *The Mysterious Affair at Styles* and compare it to a published translation. 20 participants were eye-tracked while reading the entire novel, alternating MT or HT every 25%. They find lower readability of MT compared to the published translation.

Guerberof-Arenas and Toral (2020; 2024) consider the literary reception of MT, PE, HT and ST from a creativity angle. Their reception studies look at reader responses to the different modalities using narrative engagement, enjoyment, and translation reception scales. They find that reading experience is not significantly different in HT and PE, but MT scored significantly lower on the three variables. However, results differ per text and language; in Dutch, for example, ST scores higher than PE and HT, which was not the case for Catalan readers. Our study builds upon this study as mentioned above, see Section 3.1 for details.

3 Measuring creativity and cognitive load

This section describes the data, annotation criteria, participants, eye-tracking device, questionnaire & RTA interviews, as well as the preprocessing of the initial dataset and statistical modelling. Combining quantitative and qualitative methods allows us to triangulate the data to understand the complex interactions of variables, such as creativity and cognitive load, to answer our RQs.

3.1 Content

The methodology of this study borrows the open dataset containing the original and translated texts, the annotations, and the questionnaire (on engagement, enjoyment and reception) from Guerberof-Arenas and Toral (2024).

The text, the science-fiction short story "2BR02B" (1962) by Kurt Vonnegut, contains 123 paragraphs, 234 sentences and 2548 words. Two professional translators created the HT and PE versions: to ensure readers would not rate a text higher due to translator preference, each translator did half of the text without MT (HT) and the other half with MT (PE). The MT was created by a customized NMT engine trained on literary texts, based on transformer architecture (Toral et al., 2023).

The ST was annotated by two professional translators. They identified 185 UCPs in the English ST. Subsequently, two professional reviewers annotated the Dutch TTs for creativity and errors. The UCPs in the Dutch TTs were annotated as either CS, Reproduction (no CS), NAs (too many errors for classification) or omissions (UCP is omitted entirely). Errors were classified and annotated using the harmonised DQF-MQM Framework,[2] which categories the type of error and its severity. For more details on these annotations, see Guerberof-

[2] https://themqm.org/error-types-2/typology/

Arenas and Toral (2022).The total number of CSs, Reproductions, and errors is shown in Table 1, including a creativity index (CI), which combines CSs and errors according to the following formula (in Guerberof-Arenas and Toral (2022)):

$$\text{CI} = (\frac{\#CS}{\#UCPs} - \frac{\#ErrorPoints - \#Kudos}{\#Words\_in\_ST}) * 100$$

Table 1 shows that for the texts we are using in this experiment, the CI for HT is the highest followed by PE and lastly by MT.

Creativity	**HT**	**PE**	**MT**
Creative shifts	79	63	26
Reproduction	105	122	143
Errors	75	221	528
Creativity Index	**41**	**25**	**-7**

Table 1: Results of the error and UCP annotation.

3.2 Participants

Eight participants (six women and two men) were recruited who voluntarily signed up using a Google form from a flyer distributed throughout universities and were paid €20 after completing the experiment. The criteria for selecting participants were to be native Dutch speakers (1) and frequent readers (2), reading at least one book per month. Five had a master's degree and three had a BA or were in the process of graduating. Participants were randomly assigned to one of the four modalities to read, resulting in two participants per modality. The experiment was reviewed by the Ethics Committee at the University of Groningen, and participants gave their written informed consent.

3.3 Eye-tracking

To gather data on cognitive load, we employ an EyeLink Portable Duo eye tracker (SR Research), combined with a 27-inch monitor with a resolution of 1024 x 768 pixels. Participants used a headstand for optimal sampling rate (2000Hz), with the headset set at a 55 cm distance to the eye tracker. The experiment was set up using EyeLink's Experiment Builder and carried out at the EyeLab of Groningen between May 1st and May 15, 2024. Participants were calibrated and validated using a nine-point grid, and participants were recalibrated if the calibration or validation was poor or when the deviations in validation were above 0.5. Interest areas were automatically created by Experiment Builder based on word boundaries, so that each word was a separate AOI. The text was presented on the monitor, and the font (Arial) and font-size (35) were chosen for readability (Minakata and Beier, 2021; Masulli et al., 2018). Each screen contained about 10 lines of text and participants could move to the next page themselves (by clicking or pressing a random key), which would be preceded every time by a drift check in the upper-left corner—the same place the first word of the new page would appear. A pause was included in the middle of the story, after about 1310 words which tended to be about 15 minutes of reading, where the original text also had a separating dinkus. After the pause, calibration and validation were repeated. There were no time constraints for the participants.

To answer our four research questions, we focused on five dependent variables: total fixation duration (TFD, duration of all fixations), first-pass time (FPT, duration before word is first exited), regression path (RP, duration before word is first exited to the right), fixation count (FC, number of fixations) and regression count (RC, number of regressions, which are fixations from words that come after the word in question). Previous eye-tracking studies have shown the relevance of these variables for measuring cognitive load (Skaramagkas et al., 2023), and for translation specifically (Vanroy et al., 2022). However, it is not always clear which specific measures will be the most appropriate for our RQs. Regression is, for instance, often associated with confusion, but also with increased comprehension and skim reading (Southwell et al., 2020). Nevertheless, TFD is often seen as a general indication of cognitive load, FPT for immediate and semantic understanding, with RP, FC and RC for contextual understanding. As we are interested in cognitive load, our main variable of interest is TFD. Due to space constrains, the results for the other dependent variables are shown in Appendix C.

3.4 Questionnaire

Although our main focus was on cognitive load, we also wanted to gather the demographics and reading habits of the participants, as well as their engagement, enjoyment, and reception of translation. Therefore, we also used the questionnaire from the previous on-line experiment (Guerberof-Arenas and Toral, 2024). In a computer in the lab, participants completed firstly the sections on demographics and reading patterns, and secondly, after the eye-tracking experiment, participants completed

the comprehension (10 items), narrative engagement (15 items), enjoyment (3 items) and translation reception (9 items) parts. The comprehension questions were multiple-choice, while the other three used a 7-point Likert scale.[3]

3.5 Retrospective think-aloud interview

Immediately after finishing the eye-tracking experiment and completing the questionnaires, an RTA protocol took place to triangulate the data and better understand our eye-tracking results. Participants were prompted with visualisations of their gaze during reading, but were free to discuss or mention any aspect of the story. Visualisations were created automatically with Data Viewer, using its Trial Play Back Animation feature, which shows a participant's gaze as it moves over the text in real-time, see Figure 2. This allows participants not only to comment on the text but also by seeing their eye movements clarify things they seemed to pause at or go back to. The interviews were conducted by one of the researchers in English for processing ease, except in three instances where the participants preferred to speak in Dutch; these are translated for analysis by the same researcher.

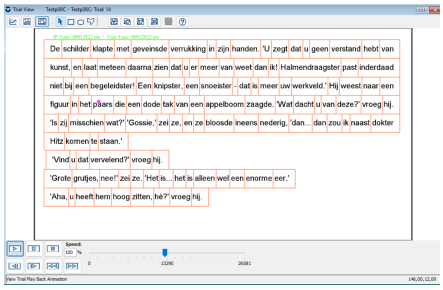


Figure 2: the Trial Play Back Animation feature in which the (moving) purple dot replays the gaze movements in real time.

3.6 Dataset and preprocessing

The data was processed using Data Viewer; 20624 observations were generated corresponding to each word in the texts, the default AOIs. These observations were further classified using the existing annotations, e.g. if a word was part of CS or Reproduction or if it was part of an error. This is our dataset I, i.e. containing data per word. Since we wanted to compare the eye-tracking data in the different translated texts for each UCP in the ST, we created a second dataset, (n=3618). In dataset II each observation is a segment either containing a UCP or without a UCP. In this way, we could compare the translation solutions for the 185 UCPs in the three translation modalities, illustrated in Figure 3. As these segments have different numbers of words, we normalised the dependent variables from the eye-tracker according to words per segment. We primarily used this dataset II for our analyses as this dataset was better suited to answer our RQs which deal with UCPs, CSs and Reproductions. We use dataset I (per word) to check word frequency and descriptive results, see Appendix C.

Figure 3: Trial of PE version, divided in segments: entire sentences or phrases. UCPs are yellow (CS) or purple (Rep.), and non-UCP segments are green.

3.7 Statistical modelling for eye-tracking data

For the statistical analysis of our eye-tracking data, we used a regression model to analyse the main effects and the interactions of our independent variables Modality (HT, PE, MT and ST), Creativity (CSs, Reproductions, omissions and NAs) and Errors (with or without errors). After analysing the data, we fitted a generalized additive mixed effect model (GAM model), using scaled t-distribution. We decided on this model as our data was not normally distributed, even after log-transformations. This was partly due to zero measurements (sections without any fixations), so we only use the data containing fixations to create the model and analyse the zeros independently.[4] We used the eye-tracking measures (TFD, FPT, RP, FC, RC) as dependent variables, with Modality, Creativity and Error as independent variables. These were contrast coded, as well as contrasting HT against PE specifically to study the difference between those two in more detail, as previous studies by Guerberof-Arenas and Toral (2020; 2024) showed little difference in reception between them. The random effects were the participants and the UCPs.

[3]Questionnaire can be found in Appendix B and on GitHub.

[4]The analysis of the zero measurement can be found in Appendix C.

The model's explained variance was about 40% for the dependent variable TFD (R^2 = 0.379), but only 28% (R^2 = 0.241) and 27% (R^2 = 0.241) for FPT and RP, while FC and RC did not meet the required assumptions for this model. Therefore, we performed non-parametric tests for those four of the eye-tracking measures (FPT, RP, FC & RC).

3.8 Analysing frequency

We also wanted to check the effect of word frequency on our data, since research has shown a strong inverse link between the frequency of words and cognitive load (Schilling et al., 1998; Holmqvist, 2011; Walker, 2021), and UCPs contain words or expressions that are less frequent. We extracted the word frequencies for all Dutch words, using the *wordfreq* Python library (Speer, 2022). This library is based on the Exquisite Corpus, which is a multilingual corpus compiled of eight different domains of text. The 'best' wordlist, available for Dutch, includes words that appear at least once per 100 million words, making it a reliable corpus. For this particular analysis, dataset I was used because word frequencies are calculated on a word-level.

4 Cognitive load, creativity and reading experience

In this section, we discuss the results from the questionnaires, eye-tracking device and RTAs that serve to answer our four RQs.

4.1 Self-reported user experience

Table 2 shows the results from the questionnaire regarding participants' demographic information and reading habits, which shows similar patterns across the participants.

Modality	n	Age	Reading habit
PE	2	18 - 24	4.5
MT	2	18 - 34	4
HT	2	18 - 24	4.5
ST	2	25 - 34	4.5

Table 2: Participants' age and reading habits from the participants (1 = Never, 2 = Once per 3 months, 3 = Once per month, 4 = Multiple times a week, 5 = Daily).

Table 3 shows the results from the questionnaire for comprehension, narrative engagement, enjoyment and translation reception from the 8 participants. Participants rated MT the lowest across all scales. HT and PE are rated higher than MT, with HT scoring higher than PE on narrative engagement and enjoyment, but not on translation reception. There are two interesting results: ST scores lower than both HT and PE on narrative engagement and reception, although not in enjoyment; and MT has the highest mean score for comprehension. Despite MT scoring the highest in the multiple-choice comprehension, the RTAs show that participants did not enjoy reading MT and reported struggling to understand the narrative (see Section 4.3). This could mean that readers in MT understand the basic details of the story so they can respond to basic questions and that they compensate using the context when they do not understand certain elements in MT; this strategy of compensation to understand MT has been reported in previous studies (Guerberof-Arenas and Moorkens, 2023).

Mod.	Category	n	Mean	SD	Med.	Min
HT	Comp.	2	8.5	0.71	8.5	8
	Eng.	30	5.37	1.35	6	2
	Enj.	6	4.67	1.21	4.5	3
	T.R.	18	4.39	1.42	5	2
MT	Comp.	2	9.5	0.71	9.5	9
	Eng.	30	3.63	1.47	4	1
	Enj.	6	2.17	1.47	2	1
	T.R.	18	2.06	1.11	2	1
PE	Comp.	2	6.5	0.71	6.5	6
	Eng.	30	5.13	1.48	5.5	2
	Enj.	6	4.17	0.75	4	3
	T.R.	18	5.06	1.06	5	3
ST	Comp.	2	9	1.41	9	8
	Eng.	30	4.20	1.58	5	1
	Enj.	6	4.67	0.82	4.5	4
	T.R.	18	4.22	1.63	4	2

Table 3: Results of the questionnaire per modality and scale (comprehension (10 multiple-choice questions), narrative engagement, enjoyment and translation reception (each 7-point Likert scale).

These results correspond moderately with the results from Guerberof-Arenas and Toral (2024): there, MT scored lowest, and ST ranked highest for narrative engagement and enjoyment. We have identified two potential causes for the difference in this second experiment: the most obvious one is that here we only had two participants per modality, which does not allow for generalization; the other reason could be that in our study, participants read the text in a lab using an eye-tracker, which could indicate higher levels of attention as opposed to reading online at home as in the original experiment.

4.2 Cognitive load

4.2.1 Descriptive results

Table 4 and Figure 4 show mean TFDs according to the variables Modality, Creativity and Errors, for dataset II (values normalised according to number of words per unit). We use TFD as this is often seen as general and overall indication of cognitive load. For Modality, ST has the highest mean duration, followed by HT, MT and then PE. This result might seem surprising: MT has the most errors (see Table 1) and previous studies have shown that errors in translated texts lead to an increase in cognitive load (Kasperavičienė et al., 2020; Stymne et al., 2012). A potential cause could be the scale of text looked at: previous studies analyse cognitive load for individual sentences with and without errors, whereas we look at the entire text. For example, Colman et al. (2022), who also looked at an entire text instead of individual sentences, also found no significant effects for modality. Furthermore, literary reading studies have shown that immersivity and engagement also increase cognitive loads in literary texts; thus, a higher quality in the literary translation–as in HT and PE–could explain a higher cognitive load and hence a higher TFD value.

TFD (ms.)		n	Mean (SD)	Median	Min	Max
Modality	HT	918	297 (260)	234	0	4042
	MT	896	219 (161)	191	0	1974
	PE	880	193 (171)	158	0	3329
	ST	924	311 (221)	248	0	1386
Creativity	CS	360	296 (251)	234	0	1840
	Rep.	728	288 (266)	222	0	4042
	Not	1606	201 (152)	175	0	3329
	UCP*	370	326 (195)	294	0	880
	Not*	554	302 (238)	230	0	1386
Errors	Yes	692	239 (171)	199	0	1974
	No	202	237 (219)	176	0	4042

Table 4: Overview of the eye-tracking data for TFD on each independent variable. UCP* and Not* refer to creative potential in ST rather than in translation.

If we compare the TFD according to creativity, we find that CSs and Reproductions (UCPs) have a higher mean than non-UCP segments, indicating that UCPs have a higher cognitive load than those segments without UCPs, but CSs do not show a difference in cognitive load compared to Reproductions. In the ST, UCPs also have a higher mean TFD compared to non-UCPs–included in the table with asterisks–perhaps due to the foregrounded elements (Jacobs, 2015; Müller et al., 2017). For Errors, as expected, units with errors have a higher mean TFD than those without, although this difference is minimal, as the box plot also shows. Moreover, standard deviations are high indicating

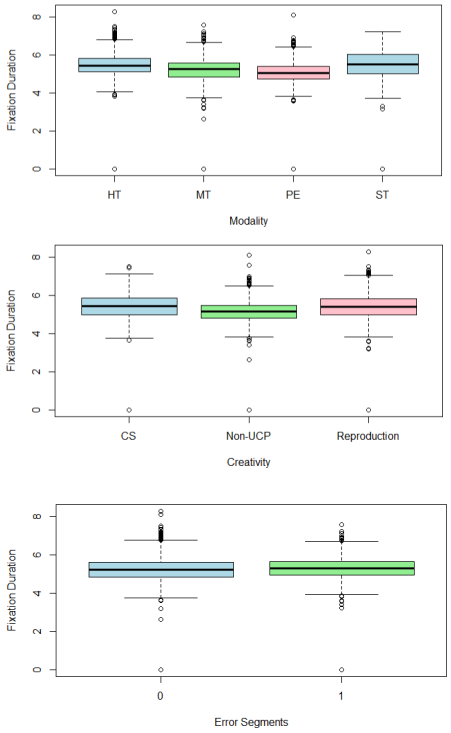


Figure 4: Box plots for the eye-tracking data (dataset II), with TFD by Modality, Creativity and Errors.

high variability across the data. This is expected due to the small number of participants and the fine-grained nature of our experiment–measuring cognitive load on word-level creates much variance, even within participants.

4.2.2 GAM analysis: main effects and interactions for TFD

Effects	Levels	Mean	SD	p-value
Intercept		1.674	0.0164	N/A
Modality	HT	0.0748	0.0396	0.059
	MT	-0.0230	0.0396	0.743
	PE	-0.0489	0.0688	0.477
	HT (v.PE)	0.1237	0.0795	0.120
Creativity	CS	0.0356	0.0084	$2.6x10^{-5***}$
	Rep.	0.0569	0.0069	$2.5x10^{-16***}$
Errors	Yes	0.0009	0.0054	0.867
Interactions between modality & creativity	HT : CS	-0.0069	0.0140	0.621
	MT : CS	-0.0452	0.0173	0.009^{**}
	PE : CS	0.0834	0.0334	0.012^{*}
	HT (v.PE) : CS	0.0765	0.0378	0.042^{*}
	HT : Rep	0.0299	0.0135	0.027^{*}
	MT : Rep	-0.0074	0.0102	0.470
	PE : Rep	0.0447	0.0175	0.011^{*}
	HT (v.PE) : Rep	0.0747	0.0237	0.001^{*}

Table 5: Main effects and relevant interaction effects from the GAM model on TFD (log-transformed duration data in ms.), ***p <.001, **p <.01, *p <.05

The results for the GAM analysis are partially shown in Table 5, including the main effects and

relevant interactions; the full table is in Appendix C. The only significant main effect is Creativity–both CSs and Reproductions. This indicates that there is an increase in cognitive load in UCPs overall (RQ1), although no clear difference between CSs and Reproductions (RQ3), as we saw for the descriptive statistics, too. There are no significant results for the independent variable Modality or Errors (RQ4). Effect sizes are less meaningful in this setting, as the GAM model is non-linear and uses log-transformations and a log-link; for an indication of differences between the levels of the separate independent variables; however, the mean TFD reported in Table 4 illustrates this effect.

There are also significant values in the interaction between Modality and Creativity (RQ2). MT has a negative effect in both interactions, significantly so for CS. So, although CSs have an increased cognitive load overall, this effect is lessened for readers in MT. We see the opposite for PE and HT, where the effect for CSs is increased for PE, and the effect of Reproductions is increased for PE and HT. Furthermore, comparing HT and PE directly reveals that in HT readers exert significantly more cognitive load on CSs and Reproductions than in PE. Readers thus seem to process CSs and Reproductions differently in different modalities, and this seems to indicate that HT has a higher level of cognition and attention in UCPs (CSs and Reproductions) (RQ2). There were no significant interaction effects for Errors with Modality or with Creativity (RQ4).

4.2.3 Frequency analysis

Lastly, we checked the effect of word frequency on our data. We correlated the word frequencies with TFD, using Spearman's correlation after checking assumptions. Results show a low negative correlation between the variables Word frequency and TFD (ρ = -0.30), although significant ($p < 0.005$). The correlations between word frequency and the other dependent variables show similar results, see Appendix C. Although an effect of frequency was expected, the low correlation here shows our reading measures are likely influenced by other factors than frequency alone.

We also created a scatter plot for the correlation, with colour-coding for Creativity to see if words assigned certain Creativity codes (CS, Rep., non-UCP) tended to be more or less frequent, see Figure 5. Colour-coding reveals no pattern, showing no clear difference between word frequencies

Figure 5: Scatter plot of word frequency and TFD (log-transformed). Colouring indicates Creativity annotation, showing no clear trend.

for words belonging to CSs, Rep. and Non-UCPs. This suggests that words in UCPs (CS and Rep.) are not necessarily less frequent than words outside UCPs.

4.3 Retrospective reading experiences

The RTA interviews were between 10 and 25 minutes per person, with a total of 144 min. of recording. This was transcribed –translated if necessary– resulting in 17732 words of transcription. One of the researchers coded the data in three cycles of coding. First, parallel coding was used, combining coding techniques such as emotion coding, concept coding and process coding (Saldaña, 2016). The focus was on the participants' (emotional) responses, what they commented on and how. This helped to understand how readers related to the story and thus helped to interpret differences in cognitive load, and to see whether and how they reacted to elements of creativity and error. In the next cycle the initial codes were combined into 11 code groups. In the final cycle, the code groups were distributed in five main themes, see Table 6. Due to space limitations, a summary per theme is given, with a detailed analysis in Appendix D.

Theme	#Codes
Confusion came from the narrative in HT, but from language use in MT	226
Engaging with and relating to narrative elements occurred in HT, ST & PE	103
HT participants felt immersed in the story, the narrative, and the style	82
MT participants had difficulty understanding the text due to nonsensical words phrasing	175
PE participants were engaged in the narrative, but struggled with the style and characters at times	106

Table 6: Main themes with number of codes included

1. Confusion came from the narrative in HT, but

from language use in MT

All participants mentioned being confused (30x HT, 37x MT, 28x PE, 23x ST). However, the cause for confusion was different across the different groups. HT participants felt confused about narrative elements, rather than phrasing: "I was a little confused by this, but that wasn't necessarily due to the words but due to the narrative" (P06_HT). Their confusion cleared after figuring out the story more. This was similar for ST, where participants were also confused about the narrative at first but this lessened as the narrative was revealed. However, MT participants were mostly confused about words or phrases that were translated incorrectly, mentioning multiple times "[it] didn't make sense" (P02_MT) and that the text was "weird" (P07_MT). They also mentioned (incorrectly translated) UCPs which caused confusion, saying in one instance: "I was completely confused (...) No idea what they did here or were intending to do." (P07_MT).

2. Engaging with and relating to narrative elements occurred in HT, ST & PE participants

HT, ST and PE participants mentioned feeling immersed in the narrative, relating it to their own lives multiple times. Participants mentioned that the story was engaging, with a well set up moral dilemma, making them empathise with Edward's (the protagonist of the story) choices: "I really sympathised with the father" (P09_PE), "I thought it was intriguing and felt bad for that man" (P08_ST), "It was sad, but it also makes you think" (P06_HT).

3. HT participants felt immersed in the story, the narrative, and the style

HT participants reported feeling immersed in both the narrative and the style throughout ("I really got into the story" (P03_HT))–something that lacked for the other modalities. HT participants specifically appreciated the style ("That was brought across very well" (P06_HT)) and repeated how much they liked the characters. Both also specifically appreciated certain translation solutions for wordplay and metaphors that were parts of UCPs.

4. MT participants had difficulty understanding the text due to nonsensical words phrasing

Most salient for MT participant was their confusion regarding the language use. Both participants did not understand many phrases or details in the story, which made it difficult to follow the story along, as they tried reconstructing the story by working back from the words to "what they should have been" (P02_MT). Both also had to laugh multiple times due to the strangeness of the MT output. P07_MT encapsulated this saying "It almost becomes poetical how bad it is." This also caused them to skim read later parts of the text as they gave up trying to understand the text with blatant errors; this might explain the lack of effect for errors on cognitive load in the eye-tracking: although errors (in MT) impact reader experience, the expected increase in cognitive load could be nullified by skim reading.

5. PE participants were engaged in the narrative, but struggled with the style and characters

PE participants were positive about the narrative, but they disliked the style, feeling it was sometimes used incorrectly or off-putting, hampering their overall enjoyment and immersion in the story. They found the moral dilemma interesting and liked the set-up, but also commented on "awkward" phrases (P09_PE) or words used out of context. P05_PE mentioned she "realised it had to be a translation, because no Dutch person would have written it like this". They also felt character descriptions were unclear or wrong, with uncommon labels ("*broeder* just confused me" (P05_PE) and oddly used adjectives (P09_PE). This could be related to the relatively low score for comprehension in PE (see Table 3 and Appendix D for more). However, this lack of comprehension could also be due to individual differences between participants.

5 Conclusions and further research

We were seeking to test our methodology with four RQs that linked cognitive load and creativity. For ease of understanding we present here the questions and the findings, followed up by the limitations of the study and future avenues of research.

RQ1: Do readers have a higher cognitive load in units of creative potential than in other regular parts of the sentence?

There was a strong positive effect of UCPs on cognitive load, both in CSs and Reproductions. Readers thus pay attention to UCPs when reading, and judging from their comments in the RTAs, they also enjoy reading them. This link between engagement and cognitive load was also found in literary reading studies (Torres et al., 2021), specifically for foregrounded elements (Jacobs, 2015; Müller et al., 2017). Although the methodology employed is solid and the results novel, the creation of the datasets is quite arduous. We think that it might be better for this word-level experiments to look at datasets that have been purposely created for this, for example, by only looking at paragraphs with

specific UCPs where it will be easier to explore the effect of modality.

RQ2: Do readers process these units differently according to the translation modality?
Although we did not find an effect for Modality, we did see a positive effect in its interaction with Creativity when looking at TFD. Furthermore, the effect of Creativity was lessened in MT, while increased in both PE and HT. Comparing HT and PE against each other specifically, we see that in HT the effect of Creativity is the highest. This was reinforced in the RTAs, where HT participants were more positive about the text, especially its style, even explicitly appreciating certain translation solutions. PE participants liked the narrative, but thought the style fell flat at times. This differs from Whyatt et al. (2024), who found increased cognitive load for low quality sentences; however, they only looked at non-literary texts, while we looked at a literary text, and literary studies have shown a positive link between the literariness of a text and fixations (Corcoran et al., 2023; Fechino et al., 2020), so that might explain the difference. Potentially, the higher overall quality of HT (as attested in the RTAs) enhances the immersive effects of UCPs.

RQ3: Do readers process the translators' solutions differently according to the level of creativity?
We did not see any significant difference in cognitive load between CSs and Reproductions. Analysing word frequency also showed only a low negative correlation between word frequency and TFD, showing that other factors influenced TFD. There was also no clustering of CSs, Reproductions. or non-UCPs, indicating that words belonging to CSs or Reproductions are not less frequent than words that do not belong to UCPs. Again here, we think that, methodologically, experiments of this type would benefit for more focused studies at paragraph-level with chosen UCPs.

RQ4: Do errors in a segment increase the cognitive load of the reader?
As expected, we see that units with errors have a higher mean TFD than those without, although only slightly. However, we do not see any significant effect on units with or without errors. This differs from previous studies on errors, as Kasperavičienė et al. (2020) and Stymne et al. (2012) found increased TFD and FC on sections with errors. Triangulating the data with the interviews, however, allowed us to hypothesise that this is due to increased skim reading in MT especially, the modality with most errors. When these readers saw too many errors, they tended to skim certain sections more. RTAs showed that errors in MT and PE influenced reading experience –creating a lack of understanding in MT and a dislike of style in PE– but this was not mirrored in the exerted cognitive load.

This study shows both the methodological advantages of analysing reader reception through cognitive load on a word-level, and the importance of creativity for reader reception across modalities. Intuitively, we expect differences in reading experiences across modalities, given errors and creativity scores. In previous studies, PE and HT scored similarly and we initially also found no main effect on cognitive load between the two; however, detailed analysis shows that creativity makes a difference; not only that, but the difference creativity makes is increased in HT. In other words, the effect of creativity is lessened in PE compared to HT. The retelling of the participants' experience through the RTAs also showed clear differences between the modalities, a crucial aspect of this methodology.

Using MT-mediated texts in literary translation thus has an impact on readers. More research into the causes of these effects is needed to inform translation technologies, translators and the industry.

We are aware of the limitations of this pilot regarding number of participants and language pairs. However, the methodology gives us a window through which we can explore the way readers deal with creativity while reading. We learnt from this experiment that document-level analysis might not be the best match to answer our RQs, therefore our next experiment, focuses on paragraphs with selected UCPs in different modalities. This will include more participants, different genres, and LLMs, to further explore this relation between creativity and reader experience.

Acknowledgments

We would like to thank all our participants. We would also like to thank Andreas van Cranenburgh for his assistance with the frequency analysis.

Funding

The INCREC project has received funding from the European Union's Horizon Europe research and innovation programme under ERC Consolidator Grant n. 101086819.

References

Gerrit Bayer-Hohenwarter. 2009. Translation creativity: How to measure the unmeasurable. In Susanne Göpferich, Arnt Lykke Jakobsen, and Inger M. Mees, editors, *Behind the Mind: Methods, Models and Results in Translation Process Research*, pages 39–59. Samfundslitteratur, Copenhagen, DK.

Gerrit Bayer-Hohenwarter. 2010. Comparing translational creativity scores of students and professionals: Flexible problem-solving and/or fluent routine behaviour. In Inger M. Mees Susanne Göpferich, Fabio Alves, editor, *New Approaches in Translation Process Research*, pages 113–139. Samfundslitteratur, Copenhagen, DK.

Gerrit Bayer-Hohenwarter. 2011. Creative shifts as a means of measuring and promoting translation creativity. *Meta*, 56(3):663–692.

Gerrit Bayer-Hohenwarter. 2013. Triangulating translation creativity scores. In *Tracks and Treks in Translation Studies: Selected Papers from the EST Conference, Leuven 2010*, pages 63–85.

Gerrit Bayer-Hohenwarter and Paul Kussmaul. 2020. Translation, creativity and cognition. In Fabio Alves and Arnt Lykke Jakobsen, editors, *The Routledge Handbook of Translation and Cognition*, pages 310–325. Routledge, New York City, US.

Sheila Castilho and Natália Resende. 2022. MT-pese: Machine translation and post-editese. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 305–306, Ghent, Belgium. European Association for Machine Translation.

Myriam Chanceaux, Françoise Vitu, Luisa Bendahman, Simon Thorpe, and Jonathan Grainger. 2012. Word processing speed in peripheral vision measured with a saccadic choice task. *Vision research*, 56:10–19.

Toon Colman, Margot Fonteyne, Joke Daems, Nicolas Dirix, and Lieve Macken. 2022. GECO-MT: The ghent eye-tracking corpus of machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 29–38, Marseille, France. European Language Resources Association.

Rhiannon Corcoran, Christophe de Bezenac, and Philip Davis. 2023. 'Looking before and after': Can simple eye tracking patterns distinguish poetic from prosaic texts? *Frontiers in Psychology*, 14.

Joke Daems and Lieve Macken. 2019. Interactive adaptive SMT versus interactive adaptive NMT: a user experience evaluation. *Machine Translation*, 33:117–134.

Joke Daems, Paola Ruffo, and Lieve Macken. 2024. Impact of translation workflows with and without MT on textual characteristics in literary translation. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 57–64, Sheffield, United Kingdom. European Association for Machine Translation.

Elisa C Dias, Heather Sheridan, Antígona Martínez, Pejman Sehatpour, Gail Silipo, Stephanie Rohrig, Ayelet Hochman, Pamela D Butler, Matthew J Hoptman, Nadine Revheim, and Daniel C Javitt. 2021. Neurophysical oculomotor and computational modeling of impaired reading ability in schizophrenia. *Schizophrenia Bulletin*, 24(1):97–107.

Gys-Walt Egdom, Christophe Declercq, and Onno Kosters. 2024. 'can make mistakes'. prompting ChatGPT to enhance literary MT output. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 10–20, Sheffield, United Kingdom. European Association for Machine Translation.

M. Fechino, A. A. Jacobs, and J. Lüdtke. 2020. Following in Jakobson and Lévi-Strauss' footsteps: A neurocognitive poetics investigation of eye movements during the reading of Baudelaire's 'Les Chats'. *Journal of Eye Movement Research*, 13(3):1–19.

Margot Fonteyne, Arda Tezcan, and Lieve Macken. 2020. Literary machine translation under the magnifying glass: Assessing the quality of an NMT-translated detective novel on document level. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3790–3798, Marseille, France. European Language Resources Association.

Ana Guerberof-Arenas and Joss Moorkens. 2023. Ethics and Machine Translation: The end user perspective. In Helena Moniz and Carla Parra Escartín, editors, *Towards Responsible Machine Translation: Ethical and Legal Considerations in Machine Translation*, pages 113–133. Springer.

Ana Guerberof-Arenas and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience. *Translation Journal*, 9(2):255–282.

Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in translation: Machine translation as a constraint for literary texts. *Translation Space*, 11(2):184–212.

Ana Guerberof-Arenas and Antonio Toral. 2024. To be or not to be: A translation reception study of a literary text translated into Dutch and Catalan using machine translation. *Target*, 36(2):215–244.

Kenneth Holmqvist. 2011. *Eye tracking: a comprehensive guide to methods and measures*. Oxford University Press, Oxford, UK.

Arthur M. Jacobs. 2015. Neurocognitive poetics: Methods and models for investigating the neuronal and cognitive-affective bases of literature reception. *Frontiers in Human Neuroscience*, 9(186).

Ramunė Kasperavičienė, Jurgita Motiejūnienė, and Irena Patašienė. 2020. Quality assessment of machine translation output: Cognitive evaluation approach in an eye tracking experiment. *Texto Livre*, 13(2):271–285.

Dorothy Kenny and Marion Winters. 2020. Machine translation, ethics and the literary translator's voice. *Translation Spaces*, 9(1):123–149.

Jeremy Klemin. 2024. The last frontier of machine translation. *The Atlantic*.

Waltraub Kolb. 2021. 'I am a bit surprised': Literary translation and post-editing: Priming effects and engagement with the text. In *Computer-Assisted Literary Translation Conference CALT2021@Swansea, 23*, pages 53–68.

Martin Kroon. 2023. *Towards the Automatic Detection of Syntactic Differences*. Ph.D. thesis, University of Leiden.

Paul Kussmaul. 1991. Creativity in the translation process: Empirical approaches. In Kitty M. Leuven-Zwart and Ton Naaijkens, editors, *Translation Studies: The State of the Art. Proceedings of the First James S. Holmes Symposium on Translation Studies*, pages 91–101. Rodopi, Amsterdam, NL.

Paul Kussmaul. 1995. *Training the Translator*. John Benjamins Publishing, Amsterdam, NL.

Paul Kussmaul. 2000a. A cognitive framework for looking at creative mental processes. In Maeve Olohan, editor, *Intercultural Faultlines Research Models in Translation Studies: v. 1: Textual and Cognitive Aspects*, pages 59–71. Routledge, London, UK.

Paul Kussmaul. 2000b. Types of creative translating. In *Translation in Context: Selected papers from the EST Congress, Granada 1998*, pages 117–126, Amsterdam, NL. John Benjamins Publishing.

KVB Boekwerk. Anderstaligheid lezen, lenen en kopen: Themameting voor de boekenbranch [online]. 2022.

KVB Boekwerk. Verkoopcijfers 2024 [online]. 2025.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Marie-Aude Lefer. 2021. Parallel corpora. In M. Paquot and S.T. Gries, editors, *A Practical Handbook of Corpus Linguistics*. Springer.

Lieve Macken. 2024. Machine translation meets large language models: Evaluating ChatGPT's ability to automatically post-edit literary texts. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 65–81, Sheffield, United Kingdom. European Association for Machine Translation.

Lieve Macken, Bram Vanroy, Luca Desmet, and Arda Tezcan. 2022. Literary translation as a three-stage process: machine translation, post-editing and revision. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 101–110, Ghent, Belgium. European Association for Machine Translation.

Naser Al Madi, Cole S. Peterson, Bonita Sharif, and Jonathan Maletic. 2020. Can the e-z reader model predict eye movements over code? towards a model of eye movements over source code. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '20 Short Papers, New York, NY, USA. Association for Computing Machinery.

Francesco Masulli, Martina Galluccio, Christophe-Loïc Gerard, Hugo Peyre, Stefano Rovetta, and Maria Pia Bucci. 2018. Effect of different font sizes and of spaces between words on eye movement performance: An eye tracker study in dyslexic and non-dyslexic children. *Vision Research*, 153:24–29.

Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.

Katsumi Minakata and Sofie Beier. 2021. The effect of font width on eye movements during reading. *Applied Ergonomics*, 97.

Tjaša Mohar, Sara Orthaber, and Tomaž Onič. 2020. Machine translated Atwood: Utopia or dystopia? *ELOPE: English Language Overseas Perspectives and Enquiries*, 17(1):125–141.

Joss Moorkens, Antonio Tora, Sheila Castilho, and Andy Way. 2018. Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7(2):240–262.

Diane C. Mézière, Erik D. Reichle Lili Yu, Titus von der Malsburg, and Genevieve McArthur. 2023. Using eye-tracking measures to predict reading comprehension. *Reading Research Quarterly*, 58(3):425–449.

Hermann J Müller, Thomas Geyer, Franziska Günther, Jim Kacian, and Stella Pierides. 2017. Reading English-Language Haiku: Processes of meaning construction revealed by eye movements. *Journal of Eye Movement Research*, 10(1):1–33.

Keith Rayner and Eyal M Reingold. 2015. Evidence for direct cognitive control of fixation durations during reading. *Current Opinion in Behavioral Sciences*, 1:107–112. Cognitive control.

Erik D. Reichle and Eyal M. Reingold. 2013. Neurophysiological constraints on the eye-mind link. *Frontiers in Human Neuroscience*, 7(1).

Mehmet Şahin and Sabri Gürses. 2019. Would MT kill creativity in literary retranslation? In *Proceedings of the Qualities of Literary Machine Translation*, pages 26–34, Dublin, Ireland. European Association for Machine Translation.

Johnny Saldaña. 2016. *The Coding Manual for Qualitative Researchers (3E)*. SAGE, London, UK.

Hildur E. H. Schilling, Keith Rayner, and James I. Chumbley. 1998. Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory Cognition*, 26:1270—1281.

Elizabeth R. Schotter, Alexander Pollatsek, and Keith Rayner. 2017. Reading. In *Reference Module in Neuroscience and Biobehavioral Psychology*. Elsevier.

Vasileios Skaramagkas, Giorgos Giannakakis, Emmanouil Ktistakis, Dimitris Manousos, Ioannis Karatzanis, Nikolaos Tachos, Evanthia Tripoliti, Kostas Marias, Dimitrios I Fotiadis, and Manolis Tsiknakis. 2023. Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE reviews in biomedical engineering*, 16:260–277.

Jungha Son and Boyoung Kim. 2023. Translation performance from the user's perspective of large language models and neural machine translation systems. *Information*, 14(10).

Rosy Southwell, Julie Gregg, Robert Bixler, and Sidney K. D'Mello. 2020. What eye movements reveal about later comprehension of long connected texts. *Cognitive Science: A Multidisciplinary Journal*, 44(10):1–24.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Maria Stasimioti, Vilelmini Sosoni, Katia Kermanidis, and Despoina Mouratidis. 2020. Machine translation quality: A comparative evaluation of SMT, NMT and tailored-NMT outputs. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 441–450, Lisboa, Portugal. European Association for Machine Translation.

Sara Stymne, Henrik Danielsson, Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz Lillkull, and Martin Wester. 2012. Eye tracking as a tool for machine translation error analysis. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1121–1126, Istanbul, Turkey. European Language Resources Association (ELRA).

Arda Tezcan, Joke Daems, and Lieve Macken. 2019. When a 'sport' is a person and other issues for NMT of novels. In *Proceedings of the Qualities of Literary Machine Translation*, pages 40–49, Dublin, Ireland. European Association for Machine Translation.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Antonio Toral, Andreas Van Cranenburgh, and Tia Nutters. 2023. Literary-adapted machine translation in a well-resourced language pair. In Andrew Rothwell, Andy Way, and Roy Youdale, editors, *Computer-Assisted Literary Translation*, pages 27–52. Routledge, New York, US.

Antonio Toral, Martijn Wieling, and Andy Way. 2018. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, 5(9).

Débora Torres, Wagner R. Sena, Humberto A. Carmona, André A. Moreira, Hernán A. Makse, and José S. Andrade Jr. 2021. Eye-tracking as a proxy for coherence and complexity of texts. *PLOS ONE*, 16(2).

Bram Vanroy, Moritz Schaeffer, and Lieve Macken. 2022. Comparing the effect of product-based metrics on the translation process. *Frontiers in Psychology*, 12.

Callum Walker. 2021. *Eye-tracking study of equivalent effect in translation : the reader experience of literary style*. Palgrave Macmillan.

Rebecca Webster, Margot Fonteyne, Arda Tezcan, Lieve Macken, and Joke Daems. 2020. Gutenberg goes neural: Comparing features of Dutch human translations with raw neural machine translation outputs in a corpus of English literary classics. *Informatics*, 7(32).

Bogusława Whyatt, Ewa Tomczak-Łukaszewska, Olga Witczak, and Olha Lehka-Paul. 2024. Readers have to work harder to understand a badly translated text: An eye-tracking study into the effects of translation errors. *Perspectives*, pages 1–21.

Elizabeth Wonnacott, Holly S S L Joseph, James S Adelman, and Kate Nation. 2016. Is children's reading "good enough"? links between online processing and comprehension as children read syntactically ambiguous sentences. *Quarterly Journal of Experimental Psychology*, 69(5):855–879.

Alfred L. Yarbus. 1967. *Eye Movements and Vision*. Springer New York, NY.

Ran Zhang, Wei Zhao, and Steffen Eger. 2024. How good are LLMs for literary translation, really? literary translation evaluation with humans and LLMs. *Preprint*, arXiv:2410.18697.

A Sustainability Statement: Energy costs and CO2 Emission Related to Experiments

Experiments in this paper made use of already existing contents (the raw MT output had been created for the previous experiment by Guerberof-Arenas and Toral (2024) and no additional models were trained, optimised or used. Therefore, there were no energy costs or carbon dioxide emissions for computational efforts related to the creation of this paper (Lacoste et al., 2019).

B Appendix: Questionnaire

The questionnaire was created in English and then translated into Dutch. As explained above, the questionnaire has a pre-task and a post-task part. The pre-task part focuses on demographics and reading habits. The demographics included questions on gender, age, education, employment and native language. The questions on reading habits asked about how often participants read, how much they enjoy reading, in which ways they read (physical book, e-book, audiobook, tablet, laptop, etc.), in which language they read (percentage-based), which genres they prefer, how often they read in Dutch and how long they read for typically.

The post-task part of the questionnaire consisted out of four sections. The first section was comprehension and was related to details of the story and were multiple-choice. As we did not discuss the story in detail in the article, we decided not to include all questions (and potential answers) here– they can however be found on `https://github.com/KyoGerrits/To-MT-or-not-to-MT`.

The other three section were scored on a 7-point Likert scale. For **narrative engagement** the questions were:

1. At times, I struggled to understand what was happening in the story
2. My understanding of the character is unclear
3. I had a hard time recognizing the thread of the story
4. My mind wandered while reading the text
5. While reading, I found myself thinking about other things
6. I had a hard time keeping my mind on the text
7. While reading, my body was in the room, but my mind was inside the world created by the story
8. The text created a new world, and then that world suddenly disappeared when the story ended
9. At times when reading, the story world was closer to me than the real world
10. During the story, I felt sad when a main character suffered in some way.
11. The story affected me emotionally.
12. I felt sorry for some of the characters
13. While reading the story I had a clear image of what the main character looked like.
14. While reading the story I could envision the situations described
15. I could imagine what the setting of the story looked like.

For **enjoyment**:

1. Did you enjoy the text?
2. How likely is it that you would recommend the text to a friend?
3. Would you consider this text high literature?

For **translation reception**:

1. The text was easy to understand
2. The text was well-written
3. I encountered words, sentences or paragraphs that were difficult to understand (including a box to write down which ones)
4. I encountered words, sentences or paragraphs that I found very beautiful (including a box to write down which ones)
5. I noticed I was reading a translation (including a box to indicate how people noticed)
6. What did you think of the translation?
7. Would you like to read a text by the same author and translator?
8. Would you like to read a text by the same author, but by a different translator?

9. Would you like to read a text by a different author, but the same translator?

In Guerberof-Arenas and Toral (2024)'s study, the Cronbach's alpha reliability coefficient (α) was 0.85 for narrative engagement, 0.87 for enjoyment and 0.79 for translation reception. These are good scores for reliability and shows the reliability of the scales. For completeness' sake, we also calculated Cronbach's alpha with our data. We had scores respectively of 0.847, 0.813, and 0.915, of which the first two are considered good and the final one excellent.

C Appendix: Eye-tracking statistics

C.1 Overview eye-tracking results per word (dataset I)

This section includes the overview of the eye-tracking data for all the dependent variables (TFD, FPT, RP, FC & RC) according to our independent variables (Modality, Creativity and Error) in dataset I, that is, the dataset per word instead of per unit, dataset II, as was used in the main analysis.

Tables 7, 8, 9, 10, and 11 show the descriptive values for each of our dependent variables (TFD, FPT, RP, FC, and RC) per word. Comparing the table with the descriptive results for the dependent variables per unit (see Table 4 for TFD and the Tables CHECK below for the other dependent variables per unit), we see similar results as we saw in the descriptive results per unit (dataset II).

IVs	Cat.	n	Mean (SD)	Med.	Min	Max
Modality	HT	5164	257 (275)	205	0	4042
	MT	5204	192 (199)	177	0	2266
	PE	5160	177 (215)	156	0	3034
	ST	5096	316 (378)	210	0	3629
Creativity	CS	1350	259 (291)	204	0	3484
	Rep.	2948	243 (264)	199	0	4042
	Not	11230	193 (216)	171	0	2560
Errors	Yes	1704	243 (249)	198	0	3034
	No	13824	205 (232)	176	0	4042

Table 7: Overview of the eye-tracking data for TFD (in ms.) on each independent variable, per word (dataset I)

IVs	Cat.	n	Mean (SD)	Med.	Min	Max
Modality	HT	5164	158 (124)	169	0	995
	MT	5204	136 (115)	155	0	1067
	PE	5160	124 (121)	138	0	981
	ST	5096	152 (127)	158	0	1129
Creativity	CS	1350	156 (122)	168	0	803
	Rep.	2948	154 (124)	166	0	981
	Not	11230	134 (119)	150	0	1067
Errors	Yes	1704	152 (118)	165	0	803
	No	13824	138 (121)	153	0	1067

Table 8: Overview of the eye-tracking data for FPT (in ms.) on each independent variable, per word (dataset I)

IVs	Cat.	n	Mean (SD)	Med.	Min	Max
Modality	HT	5164	279 (506)	202	0	20055
	MT	5204	205 (351)	173	0	8806
	PE	5160	214 (362)	161	0	7335
	ST	5096	289 (535)	191	0	16107
Creativity	CS	1350	286 (446)	200	0	60313
	Rep.	2948	264 (400)	197	0	8330
	Not	11230	277 (552)	185	0	16107
Errors	Yes	1704	152 (118)	165	0	803
	No	13824	228 (415)	176	0	20005

Table 9: Overview of the eye-tracking data for RP (in ms.) on each independent variable, per word (dataset I)

IVs	Cat.	n	Mean (SD)	Med.	Min	Max
Modality	HT	5164	1.201 (1.161)	1	0	13
	MT	5204	0.962 (0.945)	1	0	8
	PE	5160	0.865 (0.956)	1	0	15
	ST	5096	1.540 (1.682)	1	0	16
Creativity	CS	1350	1.208 (1.240)	1	0	13
	Rep.	2948	1.148 (1.127)	1	0	15
	Not	11230	0.950 (0.975)	1	0	10
Errors	Yes	1704	1.173 (1.113)	1	0	11
	No	13824	0.990 (1.024)	1	0	15

Table 10: Overview of the eye-tracking data for FC on each independent variable, per word (dataset I)

IVs	Cat.	n	Mean (SD)	Med.	Min	Max
Modality	HT	5164	0.204 (0.475)	0	0	5
	MT	5204	0.142 (0.574)	0	0	29
	PE	5160	0.151 (0.425)	0	0	10
	ST	5096	0.165 (0.451)	0	0	6
Creativity	CS	1350	0.221 (0.554)	0	0	10
	Rep.	2948	0.184 (0.700)	0	0	29
	Not	11230	0.154 (0.418)	0	0	6
Errors	Yes	1704	0.218 (0.880)	0	0	29
	No	13824	0.159 (0.426)	0	0	6

Table 11: Overview of the eye-tracking data for RC on each independent variable, per word (dataset I)

For TFD, we see higher mean results here, although median scores overlap considerably. Furthermore, we see higher SDs here as well. This makes sense as all individual words are included in this dataset (dataset I) and dataset II is normalised per word over units, reducing variance. However, the trends remain the same here. For FPT, HT has a higher mean FPT than ST here, and the differences are less pronounced than in dataset II. This is similar to FC and RC in dataset I, where we see low measures overall and that the mean of HT is higher than that of ST. For RP, the high SDs across all conditions stand out. This could be caused by relatively long regressions if a word was not understood or new, while other words were looked at much easier and quicker. In terms of mean and median values, we still see comparable results as in the other dependent variables and as in the dataset II.

We also include the box plots for TFD for the independent variables (Modality, Creativity and Errors) for all units per word. These are shown in Figure 6. These figures show similar results as

Figure 4, with HT and ST showing slightly higher TFD values compared to MT and PE. CS and Rep also have slightly higher values than Non-UCPs. Finally, the units that have a presence of errors has narrowly higher TFD scores than those without errors. The box plots for the other dependent variables show a similar trend, with little difference between the conditions.

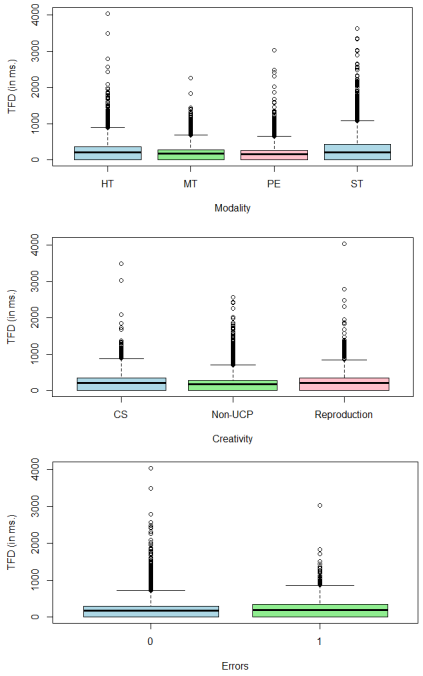


Figure 6: Box plots of TFD (in ms.) for all independent variables (Modality, Creativity and Errors), for dataset I (word-level).

C.2 Overview eye-tracking data for FPT, RP, FC, RC

This section shows the descriptive values of the eye-tracking data for our other dependent variables (FPT, RP, FC, RC) for database II. The results for TFD are in the main body, see Table 4.

IV	Cat.	n	Mean (SD)	Med.	Min	Max
Modality	HT	918	164 (71)	159	0	644
	MT	896	145 (64)	139	0	484
	PE	880	130 (81)	122	0	1644
	ST	924	158 (79)	150	0	584
Creativity	CS	359	165 (79)	161	0	674
	Rep.	728	158 (72)	152	0	644
	Not	1607	137 (72)	131	0	1644
Errors	Yes	692	151 (67)	142	0	674
	No	2002	145 (76)	138	0	1644

Table 12: Overview of the eye-tracking data for FPT (in ms., normalised for words per unit (dataset II)) on each independent variable.

Comparing the descriptive values of FPT, RP, FC, and RC in Tables 12, 13, 14 and 15 with those in Table 4 for TFD, we see similar results across the board with some small difference. For FPT, we see that mean score for HT is higher than the mean score of ST. For RP, on the other hand, the difference between ST and the other modalities is higher. For FC, the only difference is that the differences between the values are smaller than they were for TFD. For RC, what stands out are the low scores (almost all between 0.151 and 0.198, with 5 out of 9 categories with a median score of zero).

IV	Cat.	n	Mean (SD)	Med.	Min	Max
Modality	HT	918	319 (346)	241	0	4042
	MT	896	250 (261)	199	0	4322
	PE	880	232 (246)	171	0	3663
	ST	924	429 (536)	258	0	3921
Creativity	CS	359	331 (329)	247	0	2774
	Rep.	728	317 (374)	228	0	4322
	Not	1607	231 (226)	189	0	3775
Errors	Yes	692	151 (67)	142	0	674
	No	2002	264 (278)	202	0	4042

Table 13: Overview of the eye-tracking data for RP (in ms., normalised for words per unit (dataset II)) on each independent variable.

IV	Cat.	n	Mean (SD)	Med.	Min	Max
Modality	HT	918	1.363 (1.012)	1	0	10
	MT	896	1.085 (0.748)	1	0	9
	PE	880	1.002 (0.812)	0.88	0	14
	ST	924	1.604 (1.173)	1.25	0	9
Creativity	CS	359	1.389 (1.067)	1	0	7.3
	Rep.	728	1.383 (1.074)	1	0	10
	Not	1607	0.995 (0.678)	1	0	14
Errors	Yes	692	1.176 (0.796)	1	0	9
	No	2002	1.144 (0.907)	1	0	14

Table 14: Overview of the eye-tracking data for FC (normalised for words per unit (dataset II)) on each independent variable.

IV	Cat.	n	Mean (SD)	Med.	Min	Max
Modality	HT	918	0.177 (0.305)	0	0	3
	MT	896	0.152 (0.312)	0	0	4
	PE	880	0.179 (0.284)	0.10	0	4
	ST	924	0.459 (0.559)	0.33	0	7
Creativity	CS	359	0.198 (0.298)	0	0	1.67
	Rep.	728	0.198 (0.299)	0	0	1.67
	Not	1607	0.151 (0.251)	0.06	0	4
Errors	Yes	692	0.171 (0.323)	0.04	0	4
	No	2002	0.168 (0.292)	0	0	4

Table 15: Overview of the eye-tracking data for RC (normalised for words per unit (dataset II)) on each independent variable.

C.3 GAM analysis

Table 16 shows all results from the GAM analysis (whereas Table 5 in the main body only showed partial results). This includes the two-way interactions with Errors (not significant) and all three-way interactions (not relevant for the RQs).

The results for the two-way interactions with Errors are not significant, although the directions

Effects	Levels	Mean	SD	p-value
Intercept		1.674	0.0164	N/A
Modality	HT	0.0748	0.0396	0.059
	MT	-0.0230	0.0396	0.743
	PE	-0.0489	0.0688	0.477
	HT (v.PE)	0.1237	0.0795	0.120
Creativity	CS	0.0356	0.0084	$2.6x10^{-5***}$
	Rep.	0.0569	0.0069	$2.5x10^{-16***}$
Errors	Yes	0.0009	0.0054	0.867
Interactions between modality & creativity	HT : CS	-0.0069	0.0140	0.621
	MT : CS	-0.0452	0.0173	0.009^{**}
	PE : CS	0.0834	0.0334	0.012^{*}
	HT (v.PE) : CS	0.0765	0.0378	0.042^{*}
	HT : Rep	0.0299	0.0135	0.027^{*}
	MT : Rep	-0.0074	0.0102	0.470
	PE : Rep	0.0447	0.0175	0.011^{*}
	HT (v.PE) : Rep	0.0747	0.0237	0.001^{*}
Interactions between Mod. & Errors	HT : Error	-0.0064	0.0125	0.610
	MT : Error	0.0013	0.0124	0.914
	PE : Error	-0.0091	0.0232	0.696
	HT (v.PE) : Error	-0.0155	0.0278	0.579
Interact. Crea. & Errors	CS : Errors	-0.0209	0.0135	0.122
	Rep : Errors	-0.0104	0.0103	0.308
Three way interactions between Modality, Creativity & Errors	HT : CS : Error	-0.0314	0.0294	0.285
	MT : CS : Error	0.0047	0.0332	0.888
	PE : CS : Error	-0.0408	0.0629	0.517
	HT (v.PE) : CS : Error	-0.0722	0.0724	0.318
	HT : Rep : Error	0.0667	0.0281	0.018^{*}
	MT : Rep : Error	0.0183	0.0215	0.394
	PE : Rep : Error	0.0301	0.0370	0.415
	HT (v.PE) Rep : Error	0.0969	0.0497	0.051

Table 16: All main effects and interaction effects from the GAM model on TFD (log-transformed duration data in ms.), ***p <.001, **p <.01, *p <.05

are not surprising. The negative effect for HT, PE and HT compared specifically to PE reveal that generally participants spent less cognitive load on errors in these modalities, which fit our intuition that errors in MT require a higher cognitive load than errors in the other modalities; however, this value is not significant.

C.3.1 Analysis of segments without fixations

IV	Cat.	# of zeros	% of zeros
Mod.	HT	25	2.7%
	MT	17	1.9%
	PE	29	3.3%
Crea.	CS	18	5%
	Rep.	12	1.6%
	Not	41	2.6%
Err.	Yes	4	0.6%
	No	67	3.3%
Total		**71**	**2.6%**

Table 17: Frequency tables for segments without fixations (compared to the total number of segments) for each IV.

To analyse the segments with no fixations–as complement to the GAM-analysis–we created frequency tables for these segments for each independent variable, as seen in Table 17. We conducted Chi-Square Goodness-of-Fit Tests for each independent variable, but there were no significant values for Modality, Creativity or Errors. There is thus no significant effect of either Modality, Creativity or Errors when participants skipped words.

C.4 Non-parametric tests

As the dependent variables FPT, RP, FC and RC did not meet assumptions for a GAM analysis, we conducted non-parametric tests for these variables across our independent variables Modality, Creativity and Errors. Non-parametric tests only work on aggregated results–one measurement per participant, or per condition when handling repeated measures–so we calculated the means per participants and per category for each of the variables, see Table 18 for descriptive data for FPT, RP, FC and RC.

Part.	IVs	Lev.	n	FPT	RP	FC	RC
2B	Mod.	MT	2602	115.80	133.75	0.793	0.279
	Crea.	CS	87	99.07	100.73	0.558	0.135
		Rep	635	128.27	157.18	0.886	0.828
		Not	1880	112.36	127.37	0.772	0.305
	Err.	Yes	411	131.20	162.18	0.988	1.258
		No	2191	112.90	128.42	0.765	0.474
3C	Mod.	HT	2582	169.46	247.67	1.316	0.155
	Crea	CS	312	178.17	295.24	1.590	0.247
		Rep	422	187.74	289.18	1.498	0.161
		Not	1848	163.81	230.16	1.228	0.138
	Err	Yes	166	178.07	268.87	1.470	0.181
		No	2416	168.87	246.22	1.305	0.153
4D	Mod.	ST	2548	183.62	317.37	2.115	0.062
5A	Mod.	PE	2580	115.87	184.76	0.688	0.113
	Crea	CS	276	140.42	241.39	0.819	0.143
		Rep	417	128.35	217.87	0.803	0.100
		Not	1887	163.81	230.16	1.228	0.138
	Err.	Yes	275	138.84	247.28	0.864	0.148
		No	2305	113.63	177.36	0.667	0.109
6C	Mod.	HT	2582	146.80	310.73	1.086	0.254
	Crea	CS	312	169.23	413.32	1.385	0.304
		Rep	422	168.05	360.88	1.268	0.296
		Not	1848	138.16	281.96	0.995	0.236
	Err.	Yes	166	150.54	300.17	1.102	0.211
		No	2416	146.54	311.45	1.085	0.257
7B	Mod.	MT	2602	156.46	277.87	1.133	0.169
	Crea	CS	87	167.50	220.55	1.081	0.163
		Rep	635	165.85	322.83	1.294	0.200
		Not	1880	152.801	265.42	1.081	0.159
	Err	Yes	411	168.44	368.23	1.457	0.257
		No	2191	154.24	261.16	1.073	0.153
8D	Mod.	ST	2548	123.18	261.98	1.004	0.261
9A	Mod.	PE	2580	131.66	241.51	1.036	0.187
	Crea	CS	276	143.29	254.11	1.192	0.246
		Rep	417	150.37	262.95	1.189	0.197
		Not	1887	125.83	234.94	0.979	0.176
	Err	Yes	275	158.68	285.09	1.193	0.255
		No	2305	128.44	236.32	1.017	0.179

Table 18: Aggregated means per participant and per category for each of the variables, for the non-parametric tests. Including number of observations per category, mean FPT, mean RP, mean FC and mean RC.

For the independent variable Modality and the dependent variables FPT, RP, FC and RC, we conducted a series of Kruskal-Wallis H tests. However, none of the results were significant. This was some-

what surprising given the differences between the modalities observed earlier for TFD. For the independent variable Creativity and the dependent variables FPT, RP, FC and RC, we conducted Friedman's Tests (RQ2). We found significant results for the dependent variables FPT ($X^2(2) = 6.3$, p = 0.042) and FC ($X^2(2) = 6.3$, p = 0.042), but not for the others. Post-hoc comparisons, using Wilcoxon Rank Sum tests with Holm-Bonferroni correction, did not yield significant results between levels of Creativity. We wanted to analyse our independent variable Creativity further, to see whether there was any difference between UCPs (CS and Rep.), comparing CS to Reproductions specifically (and leaving out the non-UCPs). We conducted a series of Mann Whitney U tests to compare data for TFD, FPT, RP, FC and RC for CS and Reproductions, but none of these tests were significant. The analyses show that creativity overall had an effect on our participants' cognitive load, as we had also seen in the GAM analysis, further supporting a positive answer to our second research question that readers have higher cognitive load in UCP than other units.

To look at the effect of Errors, we conducted a series of Friedman's Test for the independent variable Errors for our dependent variables FPT, RP, FC and RC, but none of these were significant. So, although errors increase reading time in general, this is not significant as the GAM model also showed. We also checked for the effect of severity (none, minor and major) and type of error. Only the first was significant, specifically for FPT ($X^2(2) = 6.3$, p = 0.042) and FC ($X^2(2) = 6.3$, p = 0.042); however, here too, post-hoc Wilcoxon rank sum tests with Holm-Bonferroni correction did not reach significance.

This supports the results from the GAM analysis on TFD, with a significant result for Creativity but not for Modality and Errors. There thus seems to be increased cognitive load for UCPs (CS & Reproductions) (RQ1), but not between CS and Reproductions (RQ3) nor for errors (RQ4).

C.5 Frequency analyses

We include here the frequency analyses for FPT and RP. FC and RC are not included as these are count data and do not meet the assumptions for Pearson's or Spearman's correlation.

Figure 7 shows the scatter plot for word frequency and FPT (log-transformed). We see a similar picture as what we saw in Figure 5, with no clear trend for Creativity across word frequency.

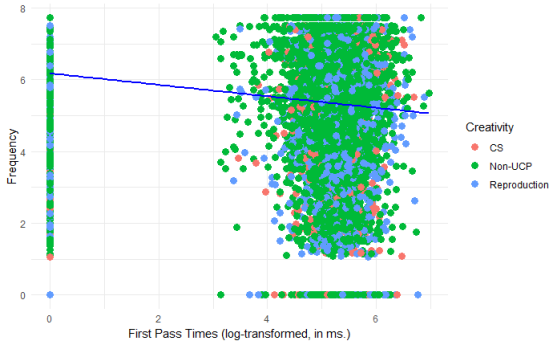


Figure 7: Scatter plot of word frequency and FPT (log-transformed). Colour indicates Creativity annotation, showing no clear trend.

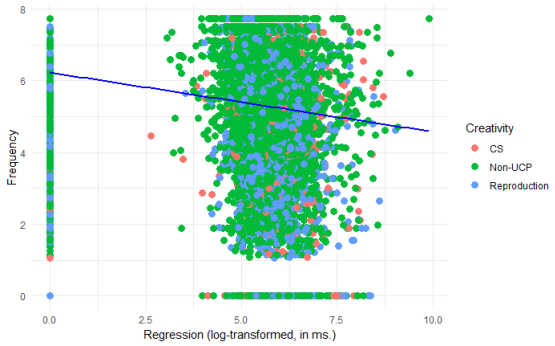


Figure 8: Scatter plot of word frequency and RP (log-transformed). Colour indicates Creativity annotation, showing no clear trend.

Spearman's correlation shows again a low negative correlation ($\rho = 0.23$)–even lower than for TFD; this too is highly significant ($p < 0.0005$). We again see that though there is some relation between FPT and word frequency this is only a low correlation, so there seems to be other factors influencing FPT as well.

Figure 8 shows the scatter plot for word frequency and RP (log-transformed). We see a similar picture as what we saw in Figure 5, with no clear trend for Creativity across word frequency. Spearman's correlation shows again a low negative correlation ($\rho = 0.29$)–similar to TFD; this too is significant ($p < 0.0005$). We again see that though there is some relation between RP and word frequency this is only a low correlation, so there seems to be other factors influencing RP.

So, for our three continuous dependent variables (TFD, FPT, and RP) we see a low negative correlation with word frequency. This means that there is a link between word frequency and cognitive load measured, but as this is low, there seems to be other factors influencing the cognitive load exerted.

D Appendix: Analysis of the retrospective think-aloud interviews (RTA)

This appendix contains the more detailed analysis of the retrospective think aloud interviews. Due to constraints of space, only a quick overview of the results per modality are discussed in the main body of the article. Here we present a more detailed analysis of the themes. The interviews were all coded by one of the researchers after which emerging themes were observed and nodes merged across consecutive coding cycles. In the end, five main themes emerged from the analysis.

1. Confusion came from the narrative in HT, but from language use in MT
2. Engaging with and relating to narrative elements occurred in HT, ST & PE
3. HT participants felt immersed in the story, the narrative, and the style
4. MT participants had difficulty understanding the text due to nonsensical words phrasing
5. PE participants were engaged in the narrative, but struggled with the style and characters at times

1. Confusion came from the narrative in HT, but from language use in MT

One of the very noticeable things is that across modalities all participants mentioned feeling confused multiple times throughout the narrative: feelings of confusion were mentioned 30 times in HT, 37 times in MT, 28 times in PE, and 23 times in ST. As expected, feelings of confusion were mentioned more in MT, but HT and PE follow closely behind. However, when looking into the reasons participants mention feeling confused, a much clearer image arises: confusion in HT refers mostly to narrative events such as the setting at the beginning and the plot twist towards the end; for MT, however, participants mentioned feeling confused largely related these feelings to words and phrases that were translated incorrectly, difficult to understand or otherwise incompatible with the context. This even made the participants laugh throughout the interview because the words "were just so weird" (P02_MT). Some clear examples in MT mentioned by both participants were *stripteasenummer* ("striptease number") for "stripling", *tripjes* (potentially diminutively morphological interpretation of the English original but also meaning "small / short trips") for "triplets", and *mooie jus* a literal translation of "good gravy". Other issues in MT included not understanding whole sentences, descriptions, or settings, due to issues with the syntax or simply too many errors. P07_MT also did not understand what was happening with the painting that one the main characters was painting throughout the story. This confusion was caused in part by the translation of "image" as *foto* ("photo") on multiple occasions while the "image" was in fact referring to a painting. P02_MT mentioned multiple times that she got confused about the syntax, such as in the sentence *Zie je hier een lijk zonder gezicht waar je me graag met je hoofd op zou willen steken?* (literally "Do you see here a (dead) body without face where you would like to put me with your head on?").

This is in clear contrast with the confusion the HT participants felt. P03_HT mentioned feeling confused about the title "2BR02B" at first, which is indeed only explained later on in the story as a phone number people can call. She also felt confused about the setting and world building, but this was due to the narrative structure of the story rather than not understanding the words and phrases used– she also got into the story really quickly after reading: on the fourth page, she mentioned "Okay, I see where this is going now" in relation to her previous confusion about the setting of the story. P06_HT mentioned being a little confused about all the nicknames of the extermination service at first, but feeling engaged in the oppressive atmosphere of the story when discovering they were all "happy-sounding nicknames for suicide machines". Both also mentioned being confused about the word *steenvruchtje* ("little stone fruit") for "drupelet", which occurred in a metaphor comparing the overcrowded world to a stone fruit-—although it is highly imaginable that people would have been confused about the original 'drupelet' as well, given that it occurs fewer than 0.01 times per million words in modern English according to the OED ("drupelet", n.1). This seems to be the case indeed as the ST participants also mentioned feeling confused about the "drupelet" the metaphor surrounding this: "I had to think about this, it takes a while to get it, it's because this [the drupelet, *red.*] got me a little confused and then, I mean what I know what he means" (P04_ST). For HT and ST then, confusion came mostly from narrative choices

or lexical choices that were similar to the original (such as the long enumeration of nicknames, which confused both P04_ST and P03_HT), while for MT confusion was caused by errors and translation issues.

2. Engaging with and relating to narrative elements occurred in HT, ST & PE

HT, ST and PE participants mentioned feeling engaged in the narrative, including the events of the story and the moral issues at play, relating it to their own lives often. For these modalities, participants mentioned that the story was interesting, with a well set up moral dilemma, making them empathise with Edward's choices. P06_HT specifically imagined how she herself would react to living in a world like that. In general, both HT participants mentioned they felt very immersed in the story, liking the characters (the second participant really loved the painter, saying he was going to be one of her new favourite characters) and describing how each character had a very distinctive style and feel to her. P08_ST mentioned he would like to read more stories by this author and that he felt very engaged in the narrative and the style: "I was also very curious to see what would happen next" and that he felt bad for the main characters too. P04_ST also said that the text was "nice to read". P09_PE mentioned really liking some of the "strong imagery" created in the story, such as the image of Leora, the word *Kattenbak* ("Catbox") for the suicide chambers and the image of *dompelen* ("dunking") people as a kind of baptism—even though he also felt that the word itself was not used completely correct: "The painter started talking about like "baptising", *dompelen* ("baptising"), and I was thinking that it was a very euphemistic description. I don't know, I thought it was interesting (...) but I also felt like it came out of nowhere and that it didn't really fit, at least not in the way it intended to". P04_ST also mentioned multiple times that she thought descriptions were chosen well and felt fitting in the story and the setting. The participants also mentioned how some of the elements of the story made them think of their own lives and experiences. P05_PE, for instance, mentioned how the description of the colour purple as "the color of grapes on Judgment Day" made her think of the art in the Galleria Borghese and the description of the character of Leora of her own mother. P09_PE mentioned relating the story to the Second World War and trying to recognise the song and creating a little melody to go with it. HT participants also related the story to their situation, with P06_HT relating the society in the story and specifically the description of the world as it was in the narrative past to the current Dutch society; she also liked the reference to Zeus in the text as a Classics' enthusiast. The other (P03_HT) started talking about a pin she herself had bought for a friend of hers which resembled Leora's pin, and towards the end, how making the appointment for the suicide chambers resembled making an appointment at the dentist. This was not the case for MT. P02_MT did not relate the text to her own life, only relating some of the in their eyes more surprising errors to text-external things: she linked the *grove vrouw* ("coarse woman") to *grove mosterd* ("coarse-grained mustard") and *ontlasten* (litt. "relieve") to peeing rather than "disposing of someone" as in the original. P07_MT did not relate any part of the story to her own life at all, only mentioning how weird things sounded or how it should have been in Dutch to reconstruct the story (e.g. "'dompelen mensen onder' ("immerse people") I didn't completely understand but it's probably about people who are dying" or "'oude eend' ("old duck") was also funny, like okay, 'old man' I'd say or 'oude lul' (litt. "old dick") or something"). Research has shown that readers who relate parts of a story to their own experiences and own frames of understanding and seeing the world feel more engaged and like a story better (Kuiken et al., 2004).

3. HT participants felt immersed in the story, the narrative, and the style

Throughout the interviews, the HT participants made it clear that they liked the story in many of its facets and felt immersed in both the narrative and the style. P06_HT kept commenting about how immersed she felt in the story, how much she liked the character of the painter, how much she empathised with Edward and how well-put the moral dilemma was. She specifically mentioned enjoying the dialogue and the "ironic and witty" banter back and forth between the painter and the nurse, describing the dialogue with Dr. Hitz from both perspectives in the story, seeing Dr. Hitz's appreciation of the system but also the "painful" decision for Edward. She also mentioned multiple comical instances in the story, such as Leora's moustache and the way she and the painter squabbled about which figure fit her best. She also described how she would feel and act from the different characters' viewpoint,

clearly immersing herself and placing herself in the story. P03_HT also engaged emotionally with the characters, describing Dr. Hitz "as being just so annoying, [which] works so well for the story". She described the story as "engrossing" and mentioned how the story kept a good balance between explaining and showing the world-building. She also pointed out the wordplay in *Duncan* and *dunken* ("dunking"), which she felt was not only very good and expressive, but also a good find on the translator's part. Both mentioned how the story made them think about the world, the story world, what they would do themselves in such a situation and whether the story world is a better world than our current world. Still there was also some confusion in the HT version, but this tended to be related to the narrative and fit general reading experiences (especially for short stories), such as confusion about world-building at the beginning and surprise at plot twists. The HT participants engaged deeply with the story, feeling immersed in the narrative and the characters, appreciating the style and the way the moral dilemma made them reconsider some of the values and situations in the world.

4. MT participants had difficulty understanding the text due to nonsensical words phrasing

Translation errors in MT led to nonsensical phrasings, which caused the participants to struggle understanding the narrative and its events. Participants mentioned that they were not sure what was happening at multiple times during the RTAs ("[I] just didn't really see what was happening here" (P02_MT) & "I couldn't follow what it said" (P07_MT)). This made it difficult for them to retain and envision the story in their minds, including which character was which, what their role and potential development was in the story and what had happened so far in the narrative: "I couldn't really connect with the characters here, she seemed very, yeah, I don't know, problematic? But yeah, it's also just like the text, you know the words and stuff, feel like there's a barrier there or something... also because I just don't really know what's actually there or what's weird or something" (P02_MT). Participants also mentioned struggling with retaining the developments of the plot in mind as they were continuously trying to reconstruct the 'correct version' of the text and events in their mind while reading. P07_MT mentioned "it was funny; you know what they are trying to say, but it does not work like that, and it is definitely not correct." P02_MT also described trying to "reconstruct" the correct version of the text in her mind, but she said that this made her feel very detached from the story and caused her to, at times, read the text cursory rather than in-depth because "[she] had no idea what was going on anyway". Rather than trying to reconstruct the narrative, she also mentioned giving up at times: "I also think here is kind of where I also started giving up? Or like, not necessarily actually giving up but more like, accepting that I wouldn't really get the thing". When discussing the metaphorical image of the drupelet, she also mentioned not even bothering to recreate the image in her mind, because she did not believe she would understand the metaphor anyway.

However, this does not mean participants hated the story. Both mentioned liking certain parts of the narrative. One of the participants felt the ending was very fitting for the story and mentioned liking the moral dilemma, describing the story as a "gripping sci-fi story" (P07_MT). Both also mentioned that they believed they would like the story a lot more in English ("[I think] I'd prefer to English original" (P02_MT). Still, it is clear that on the word and stylistic level, MT was strongly inadequate, obfuscating understanding of the text and even for those sections where the meaning could be reconstructed making readers feel detached and disengaged from the different story elements and the plot as a whole.

5. PE participants were engaged in the narrative, but struggled with the style and characters at times

PE participants liked the story overall, thought it set-up the moral dilemma really well, and enjoyed themselves while reading the story. Both participants related the situation and parts of the setting in the story to their own lives and experiences, and one of the participants specifically mentioned the "strong imagery" (P09_PE) in the story throughout. When PE participants expressed their confusion, this tended to be related to the narrative elements in a similar way as the HT participants' confusion, rather than any confusion caused by nonsensical phrasing or other (blatant) translation errors as happened for MT participants. At the same time, however, PE participants did not like the style: P09_PE, who was a little milder than the other, said that the style "did not struck [him] specifically", but liked it well enough, although he also mentioned that "some sentences seemed off, not specifically

clear why but the phrasing seems off". P05_PE was more forceful and negative about the style, saying that "[she] realised it had to be a translation, because no Dutch person would have written this like this". However, both participants found it difficult to exactly pinpoint instances in which they disliked the writing. This could be caused by the fact that there were almost no glaring errors in the text (which MT did have), but rather just a general feeling of the text not adhering to normal Dutch writing styles. The participants did mention some adjectival use that felt strange and some of the words that seemed to be out of context. The instance of *broeder* ("brother") for the nurse is discussed above, in which the chosen translation is not so much incorrect, but rather uncommon and more commonly used in other contexts (monks or in the rap and street scene colloquially). It is possible that there were more of such instances, which were less conspicuous but influenced the reading experience.

One of the other surprising things that happened with both PE participants is that they confused multiple characters. P09_PE confused Edward (the father) with the painter, while P05_PE confused Edward with Dr. Hitz. It is true that the story does not have a clear main character, with all three playing an important role in different parts of the story, but it is noticeable that both participants had issues with keeping the characters apart. This was also not just a brief confusion of characters, but both participants only realised their error during the questionnaire when the multiple-choice options included all characters. It is a little unclear what caused this confusion. Both participants mentioned that characters' motivations were not always clear, although P09_PE said he liked the characterisation overall. P05_PE was more critical about the characterisation, feeling that it was done "rather poorly", with characters' emotions shifting immensely without any explanation or emotions she could not place in general, also mentioning that she "couldn't really connect to the characters". P09_PE did comment that the adjectival use was weird throughout the story, especially pointing to the adjectives that were used to describe characters, such as *een grimmige oude man* ("a grim old man") *een grove (. . .) vrouw* ("a coarse woman"), and P05_PE also mentioned feeling confused about the description of the hospital brother as *broeder*, which in Dutch is acceptable but not very commonly used. Both also mentioned that they felt shifts in the story were very sudden and that the different sections were not well connected.

Lastly, PE participants seemed more confused (and for a longer period of time) than HT participants: like the HT participants, both PE participants mentioned feeling confused about the image of the drupelet; however, PE participants seemed to understand the imagery only later during the RTAs, while HT participants said they understood it almost directly when reading the text for the first time. P05_PE also mentioned not fully understanding the ending: "Everyone dies and then you have the painter, and he continues to paint or something?"; although this also relates to the narrative level as the confusion in HT did, it seemed that in PE these confusing elements were not always solved (as they were in HT), which left PE readers with a lower appreciation (as shown in the RTAs) and potentially comprehension (as shown in the questionnaire) than HT readers had for these narrative elements.

Interestingly, it were also the PE participants who had the lowest score for comprehension in the questionnaire (a mean score of 6.5, see Table 3)–this relative low comprehension could be linked to the confusion the participants mentioned in the RTAs. A potential cause for this confusion and lower comprehension in PE are the lack of connections and particles in the PE version. PE participants mentioned that they felt shifts in the story were very sudden and that the different sections were not well connected: "it [the narrative, *red.*] seemed to jump around in like the setting and characters and like the shifts from one thing to the next were a bit inconclusive, or random". These sudden jumps and shifts could be caused by the lack of connectors and particles in PE, which are typical of Dutch language and studies have shown PE struggles with these at times (Kroon, 2023; Lefer, 2021). This could cause the lack of cohesion felt by the PE participants which in turn could potentially explain the lower comprehension of PE participants throughout. However, it could also be that the specific PE participants just struggled more with the text or that for these two participants the experimental conditions (such as reading from a computer while resting their head in a headset) had more impact on their general reading experience.

Translation Analytics for Freelancers: I. Introduction, Data Preparation, Baseline Evaluations

Yuri Balashov
University of Georgia
Athens, Georgia, USA
yuri@uga.edu

Alex Balashov
Evariste Systems, LLC
Athens, Georgia, USA
abalashov@evaristesys.com

Shiho Fukuda Koski
SFK Language Solutions
Rochester, New York, USA

Abstract

This is the first in a series of papers exploring the rapidly expanding new opportunities arising from recent progress in language technologies for individual translators and language service providers with modest resources. The advent of advanced neural machine translation systems, large language models, and their integration into workflows via computer-assisted translation tools and translation management systems have reshaped the translation landscape. These advancements enable not only translation but also quality evaluation, error spotting, glossary generation, and adaptation to domain-specific needs, creating new technical opportunities for freelancers. In this series, we aim to empower translators with actionable methods to harness these advancements. Our approach emphasizes Translation Analytics, a suite of evaluation techniques traditionally reserved for large-scale industry applications but now becoming increasingly available for smaller-scale users. This first paper introduces a practical framework for adapting automatic evaluation metrics—such as BLEU, chrF, TER, and COMET—to freelancers' needs. We illustrate the potential of these metrics using a trilingual corpus derived from a real-world project in the medical domain and provide statistical analysis correlating human evaluations with automatic scores. Our findings emphasize the importance of proactive engagement with emerging technologies to not only adapt but thrive in the evolving professional environment.[1]

1 Introduction

This is the first in a series of papers exploring the rapidly expanding new opportunities arising from recent progress in language technologies for individual translators and language service providers (LSPs) with modest resources.

1.1 Background and related work

Many translators use MT output in their workflow. In fact, MTPE (machine translation post-editing) has become the default *modus operandi* in the industry (Pérez, 2024) and is seamlessly integrated into computer-assisted translation (CAT) tools and translation management systems (TMS). (For a recent review, see Moorkens et al., 2025, Ch. 8.) Most CAT tools can now send real-time queries over the Internet (widely referred to as "API calls") to any number of generally available neural machine translation (NMT) engines or MT aggregators and present the retrieved translation suggestions to the users for their consideration, alongside translation memory (TM) matches.

The advent of large language models (LLM) made the work environment of a typical freelancer more complex because, among other things, LLMs can translate, demonstrating performance competitive with that of dedicated NMT engines for some language pairs and domains (Castilho et al., 2023; Fernandes et al., 2023; Garcia et al., 2023; Hendy et al., 2023; Peng et al., 2023; Wang et al., 2023; Zhang et al., 2023; Peters and Martins, 2024; Li et al., 2024b; Li et al., 2024a; Lyu et al., 2024; Zhu et al., 2024). Even more importantly, with the right prompting, they can perform increasingly more sophisticated and advanced operations including, but not limited to:

- Evaluating the quality of translation output, including their own (Kocmi and Federmann, 2023; Lu et al., 2024), with or without reference translations.
- Spotting and categorizing translation errors and suggesting corrections (Berger et al., 2024; Feng et al., 2024).

[1]Data: https://github.com/YuriBalashov/reeve-corpus. Code: https://github.com/abalashov/llm-translation-testbed/.

Proceedings of Machine Translation Summit XX Volume 1, pages 538–565, June 23-27, 2025

- Automatic post-editing of raw MT output, including their own (Raunak et al., 2023; Ki and Carpuat, 2024; Alves et al., 2024; Rei et al., 2024).
- Adapting translation output:
  - to the required terminology (Ghazvininejad et al., 2023; Rios, 2024);
  - to a given domain (e.g. medical, legal, IT, aerospace engineering, etc.) (Sia and Duh, 2023; Zheng et al., 2024);
  - to existing translation memories and other project-, client- or domain-specific instructions and reference materials, often outperforming in these respects more traditional approaches earlier implemented in NMT systems (Moslem et al., 2023; Moslem, 2024; Vieira et al., 2024).
- Generating mono- and bilingual glossaries of special terms from pairs of source and target documents (Ding et al., 2025; Halpern, 2025)
- Improving the quality of translation in low-resource directions (e.g. DE-HI) by following a COT-style ("chain-of-thought," Wei et al., 2023) prompt which explicitly requires them to pivot ("Translate this sentence from DE to EN first; then translate the EN output to HI"); see, in particular, Jiao et al., 2023.
- Following, with benefit, a human translation workflow (Chen et al., 2024; He et al., 2024) by engaging LLMs in a multi-turn interaction involving pre-translation research, drafting, refining, and proofreading (Briakou et al., 2024).

The possibilities in this area are virtually unlimited. Tech giants, larger LSPs, and MT aggregators are losing no time experimenting with these and other approaches in the context of massive localization workflows, with the goal of reducing the role of the proverbial "human expert in the loop" to the very minimum (see, e.g., Intento, Inc., 2024; Zekpa and Peter, 2025; RWS Group, 2025). CAT and TMS developers are hurrying to incorporate the latest LLM-powered features into their systems (e.g. memoQ, 2025: Bureau Works, 2025). New dedicated LLM-based applications are being offered to human translators,[2] sometimes premised on the assumption that translation memory is a depreciating asset.

[2]E.g. CotranslatorAI.

1.2 Our goals in this series of papers

There is no doubt that these trends will continue to shape the future of translation, human and machine, and will introduce numerous new and unforeseen changes to the fundamental nature of our work. Freelance translators, like everyone else, are adapting to the ongoing changes brought about by the latest developments in AI to the best of their ability. While this adaptation is crucial to the future of the profession, we submit that to get ahead of the curve, a more proactive approach is required.

Linguistic expertise has always been a distinctive mark of excellence in human translation work. However, freelancers are asked to perform other tasks such as sentence alignment, TM clean-up or glossary creation. In our own experience as translators, these tasks are growing in demand, which is consistent with anecdotal evidence from our colleagues and recent industry reports which emphasize "an increasing need for human translators to occupy new roles" (Crangasu, 2025), such as "AI Content Strategy," "Big Data Curation," or "QA Automation" (Da Fieno Delucchi et al., 2025). See also Slator, 2024; Al-Batineh and Al Tenaijy, 2024.

Freelancers are also increasingly asked to offer their advice on the quality of project- or domain-specific linguistic resources such as TMs or termbases (TB). Use cases include "a company looking to improve its AI translations," a task that requires "experienced translators to pour through large volumes of the translated text" (Crangasu, 2025). A request to compare the relative quality of several candidate TMs for a given project is another good example of a task that would benefit from a novel combination of linguistic and technical knowledge. In some cases, pairwise automatic scoring of one TM against another, used as a reference, may be a good first step in the process. We believe that developing new technical skills proactively would make us better prepared for the upcoming challenges. To put it in slogan form, this could make a difference between the "AI is taking our jobs" and "AI is creating new opportunities for us" standpoints pervading much of the current discourse about AI.

Needless to say, many translators already have sophisticated technical capabilities. We think, however, that *Translation Analytics*—an umbrella category we shall use to refer to a variety of methods for the evaluation of the quality of translation-related linguistic assets—have not been deployed by free-

lancers to its full capacity. In fact, for most of them, 'Translation Analytics' may be synonymous with pre-translation analysis performed by CAT tools to generate the statistics for fuzzy TM matches at the start of a new project—for pricing, time planning, and other business purposes. Translation Analytics, however, are much broader in scope. We think of them as including, but not limited to:

- Human evaluation methods ranging from linear scoring to Multidimensional Quality Metrics (Lommel et al., 2013; Freitag et al., 2021; Knowles and Lo, 2024; Lommel et al., 2024).
- Automatic evaluation metrics, such as BLEU (Papineni et al., 2002), chrF (Popović, 2015), TER (Snover et al., 2006), and COMET (Rei et al., 2020).
- Any number of *ad hoc* tools and methods for statistical analysis and quality estimation that may be developed for a given project and tailored to its specific demands.

Our main goal in this series of papers is to explore the full potential of Translation Analytics in the context of a typical freelancer workflow. We aim to empower fellow translators with new methods that would allow them to add value to their services at the time of big changes and to gain control of the processes that usually happen "under the hood." We also hope this will stimulate developers of CAT/TM tools and TMS systems to incorporate some of the analytic methods we describe in this series of papers into their products.

In the end, freelancers should be able to implement many of the sophisticated operations mentioned in Section 1.1 above, in their local translation environment, with practical, theoretical, and strategic benefits. Instead of contributing the last, indispensable but increasingly small, bit of human expertise to the proverbial "loop" the translator can get back into the driver's seat by learning a small number of new technical skills.

1.3 Our goals in this first article

In the first article in this series we focus on adapting automatic evaluation metrics to the needs and work environment of individual translators and smaller LSPs who may want to take their technical capabilities to the next level.

Automatic evaluation of MT quality has been a prominent focus in the industry for years. Traditional metrics like BLEU (Papineni et al., 2002), chrF (Popović, 2015), and TER (Snover et al., 2006) assess the output of MT systems by comparing it to reference translations, ideally created by skilled human translators. These comparisons rely on word or character-level string matching. Newer metrics such as COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020), and BERTScore (Zhang et al., 2020) evaluate translations within the semantic space of neural networks. This approach is less reliant on specific word choices and instead prioritizes the underlying linguistic meaning.

The correlation between automatic metrics and human evaluation remains a topic of debates (for a recent overview of these debates, see Moorkens et al., 2025, Ch. 5), yet these metrics are essential in MT research and development. They enable developers to quickly compare model outputs after numerous adjustments to determine whether a particular change improves quality. Additionally, automatic metrics can monitor the training of NMT models by calculating, for instance, a BLEU score on a reserved test set after each iteration. Training can be stopped when no further improvement is observed.

Historically, automatic metrics were both technically complex and irrelevant to human translators, who depended on their linguistic expertise and manual analysis. However, with the seamless integration of MT engines into CAT tools, the vast availability of bilingual data at translators' fingertips, and recent advancements in generative AI, the landscape is evolving rapidly. Many translators now incorporate MT into their workflows and often need to choose among multiple MT engines for specialized projects, sometimes spanning tens of thousands of words. Translators frequently possess valuable bilingual resources, such as TMs and TBs from similar projects, which allow them to evaluate MT engine outputs in minutes using automatic metrics. Free online tools designed for users without programming expertise facilitate this process.[3] One such tool, MATEO (MAchine Translation Evaluation Online) (Vanroy et al., 2023), is used in our work.

To illustrate the power and practical value of such methods for individual translators, we need high-quality data—parallel documents in two or more languages. While most of industrial-scale translation quality evaluation research is based

[3] And many other processes. Thanks to free online toolkits such as adaptNMT (Lankford et al., 2023), anyone can now build, train, fine-tune, and evaluate an NMT system more or less from scratch!

on the datasets made available on Workshops on Machine Translation (WMT) benchmarks (Kocmi et al., 2024) and other shared task repositories, we take our data from a recent real-life translation project completed in summer 2024 for a non-government organization, as described below.

Our main contributions detailed in this paper are as follows:

- We present, with the client's permission, a **trilingual corpus of over 4.5K sentences in English, Russian, and Japanese in the medical domain (the Christopher & Dana Reeve Foundation Trilingual Corpus, RFTC)**, resulting from a recent human translation project completed by YB (EN-RU) and SFK (EN-JA) who are certified by the American Translators Association in their respective language pairs. We hope this corpus will be used for non-commercial research purposes by others and that it will grow both in coverage and language varieties.
- We use this corpus to develop and implement a relatively simple **approach to translation quality evaluation** which can be adapted by technically oriented translators and LSPs with modest resources to assess the quality of translation output from traditional NMT engines and LLMs in an informed way.
- We report the **BLEU**, **chrF2**, **TER**, and **COMET** scores for the translation outputs (EN-RU and EN-JA) for a slightly smaller but more challenging version of our corpus (about 3.5K English sentences) from **three popular NMT engines** (labeled MT1–MT3) and **three popular LLM models** (LLM1–LLM3), using our professional human translations for reference (see Appendix A for details).
- We adopt a simple linear 0.0–4.0 scale **modeled after academic grading** to perform preliminary **human evaluation** of **540** MT- and LLM-produced translations in each of our target languages (i.e. **1080** sentences overall).
- We report and discuss the results of our **preliminary statistical analysis** in order to determine:
  - whether the automatic scores computed for smaller non-overlapping parts of our source document (229, 1143, and 2183 sentences) correlate with each other;
  - whether sentence-level COMET scores for select segments for each output correlate with the human grades for them.

Two of the authors (YB and SFK) are ATA-certified professional translators with little or no programming experience or skills. Our perspective, therefore, fits the goals of this use case study. We should add that while we could, in principle, meet our coding needs by asking LLMs to write simple programs for our operations, coding with LLMs can be a haphazard process; the output can be very good and correctly focused on the problem, or can be mediocre and not especially applicable to what one is trying to accomplish. Stitching it all together for the purpose of a systematic study is still a task that benefits from a great deal of human expertise. Our experiments would be far from complete at this point without the tremendous help from a professional programmer on our team (AB) who took care of all the LLM-related operations, API call parallelization, streamlining, and more, as described in Section 3 below.

We adapt our discussion throughout the paper to the specific needs of individual translators and smaller LSPs. While the size of our corpus is small by MT industry standards, it is quite large for a single human translation project, and it generates statistically significant evaluation data. Furthermore, since our corpus is unlikely to have been seen and used for training or fine-tuning by generic NMT engines and popular LLMs at the time of conducting our baseline experiments, it adds new evidence for the ongoing debate about the quality and reliability of automatic quality metrics.

We believe that getting under the hood of translation quality evaluation is very important for freelancers and smaller LSPs at the time when traditional workflow models are being replaced by increasingly more sophisticated tasks requiring new technical expertise and willingness to learn more advanced methods. We submit that equipping individual translators with the additional technical capabilities described in this series of papers will help them adapt their toolkits to the rapidly changing work demands and new challenges brought about by the rocket speed development of language technologies.

The plan for the paper is as follows. Section 2 presents the Reeve Foundation Trilingual Corpus, complete with our reference translations. Section 3 describes how we obtained MT and LLM translation outputs for our source documents. In Section 4 we report and discuss the automatic metric scores for the entire corpus. In Section 5 we investigate pairwise correlations among the scores for three

smaller parts of the corpus. In Section 6 we develop our approach to human evaluation of select MT and LLM output, present its results, and discuss their statistical significance. In Section 7 we note the limitations of our study and outline plans for future work. Section 8 summarizes our findings and conclusions.

2 The Christopher & Dana Reeve Foundation Trilingual Corpus

We illustrate our Translation Analytics methods with the resources from a large translation-editing-proofreading project completed in summer 2024 for the Christopher & Dana Reeve Foundation.[4] Specifically, the Foundation's Paralysis Resource Guide is "a free comprehensive 392-page book designed to empower individuals living with and impacted by paralysis to lead healthy and fulfilling lives." A shorter (80K words) international edition of the Guide was recently translated into several languages. The Guide (referred to below as 'PRG') is a coherent structured document divided into chapters and sections, complete with a descriptive glossary of about 200 technical terms. The translation project (EN-RU and EN-JA) came in the form of IDML (InDesign Markup Language) files for separate chapters. The PDF layouts of the EN, RU, and JA versions of PRG are included in our corpus for reference.

As the first step in data preparation, we took the versions of our TMs which preserve the order of source sentences in the original full document. We removed IDML and other tags from the TMs, discarded repetitions, and produced a spreadsheet that combined the source text (EN) and our reference translations (RU, JA). Next, we performed additional cleanup operations to remove:

- leading and trailing spaces;
- bullets and other special characters at the beginning of segments;
- segments with only or mostly numbers;
- segments with only or mostly URLs;
- segments with only or mostly address lines or phone numbers.

The resulting Excel file `1-10_en-ru-ja_long.xlsx` contains 4528 segments supplied with stable ID numbers (Column A), which are used in all our experiments.

To make the translation task more challenging for MT engines and LLMs, we also decided to remove segments shorter than 6 source words from our set and generated "short" versions of the data (`1-10_en-ru-ja_short.xlsx`, etc.). The source sentence length (`Len`) is calculated in Column F.

An additional minor reduction was necessitated by the limitations MATEO imposes on the input file size ($\leq$ 1MB) for evaluation (Section 4 below). To preserve the natural order of the segments, we met this requirement by removing the last two parts of PRG ("Glossary" and "Back Cover"), which brought the segment count down to 3555 (`1-8_en-ru-ja_short.xlsx`). The resulting Excel document was used to prepare tri- and monolingual Unicode text files for our experiments.

The materials referenced above comprise the Christopher & Dana Reeve Foundation Trilingual Corpus (alternatively, the Reeve Foundation Trilingual Corpus, RFTC), complete with the PDF layouts. Additional corpus details can be found in Appendix B. With the client's permission, we make the corpus described here available for non-commercial/academic use.

3 Translation Outputs

In this section we describe how we obtained MT and LLM translation outputs for our corpus.

3.1 Technical notes on MT output

To preserve data confidentiality, we used the "Pro" versions of three popular NMT engines (labeled MT1, MT2, and MT3) to translate the entire `1-10_en_short.txt` document (3896 segments, one per line). The process was implemented as "pre-translation" in memoQ for MT1 and MT3, and was performed directly for MT2. We tracked the run-times for these operations (Table 5 below).

3.2 Technical notes on LLM output

Translation with LLMs was more complicated. API calls over the Internet must be used to interact with the major LLMs because the latter offer both the use of the model, and the significant parallel computing resources required to run it, as an integrated, metered "cloud" service. We used the Python programming language and the Python SDKs provided by major LLM vendors. We used paid subscription accounts for all LLM calls, with maximum data security/privacy settings allowed for these accounts.

[4]The authors thank the Christopher & Dana Reeve Foundation for a kind permission to use their linguistic resources in this work.

3.3 Bulk processing and LLMs

There are, in principle, a number of ways to feed a large list of sentences to major LLMs. Some of them, for example, offer API constructs for batch processing, specifically intended for non-time-critical bulk tasks. In this approach, large data sets are uploaded for the LLM provider's backend to churn through on a best-effort basis. To limit scope creep and eliminate variation in how we used different LLMs, we did not explore this option. It is also possible to submit multiple sentences with every request; this we did try, but we found the formatting characteristics of the resulting output to be too inconsistent for automatic evaluation. Therefore, the only method we evaluated was sentence-by-sentence, with one sentence per request.

It is worth taking a moment to reflect on the fact that this sentence-by-sentence approach is relatively naive, in a sense, even if it also eliminates some confounding factors. In contrast to the contextual environment of an ongoing ChatGPT conversation, in which the model keeps a running context window where prior prompts and responses reside, every one of our API requests instantiated a *de novo* context that was not informed by prior state. We did not attempt to evaluate the impact of context windows upon translation quality for two reasons: (1) the additional variables introduced would be unwieldy for the modest ambitions of this paper, and (2) some *ad hoc* experimentation did not suggest that there was much, if anything, to be gained in translation *quality* this way, and therefore it did not seem a propitious avenue for our specific aims. Still, this may be worth exploring in future research.

3.4 Prompt specificity

We found that brief and broad requests are not rewarded with as much consistency as long and specific ones. For example, when commanded to `"translate the following sentence to Russian:_____"` major LLMs would, for the most part, return the translated sentence and nothing else. However, every once in a while, the resulting sentence would contain additional verbiage: `"Here is the following sentence in Russian: _____"`.

With a more laborious prompt, which spelled out some examples of extraneous contributions unrelated to the translation of the sentence, this effect could be mostly, but not entirely obviated:

`"You are an expert translator, translating for an expert audience. Please do not provide any annotations, explanations or transliterations in your translation. Please translate the following sentence to Russian (Japanese): _____"`

Rarely, extraneous output would still appear, although the prompt was highly effective at reducing the incidence of it. (We did not specifically attempt to measure the incidence.) This is a salient consideration for any endeavor that relies on low-touch bulk translation by LLMs.

3.5 Temperature and determinism

It is well known that LLMs' output is not 100% deterministic. All LLM providers offer an API call parameter called "temperature" (T) which regulates the degree of acceptable stochastic variance in responses; higher temperatures allow more randomness, and lower values less. We set $T = 0.0$ in all of our requests across the board, but occasional variation in responses to identical prompts, while rare, was still present.

3.6 "Buggy" prompts

Upon completing all the operations with LLM and collecting all the outputs, we discovered that our optimized prompting routine concatenated the language name (i.e., 'Russian' or 'Japanese') to the prompt prefix twice:

`"You are an expert translator, translating for an expert audience. Please do not provide any annotations, explanations or transliterations in your translation. Please translate the following sentence to Russian (Japanese): Russian (Japanese): _____"`

Given the length of the prompt, we hypothesized that this did not have a significant impact on the output. But we decided to perform a safety check comparing the LLM outputs for a shorter part of PRG (`5.en_short`, 229 segments) with the above prompt (which we used in our experiments) as well as with the corrected prompt:

`"You are an expert translator, translating for an expert audience. Please do not provide any annotations, explanations or transliterations in your translation. Please translate the following sentence to Russian (Japanese): _____"`

We generated two outputs with the "bug-free" prompt to see if the differences between them due to the usual sampling (even with $T = 0.0$) in LLMs are significantly smaller than the differences between each of them and the output for the "buggy"

prompt. The results reported in Appendix C suggest that the answer is No. In terms of automatic scores, the differences among the three outputs are marginal and statistically insignificant, both for EN-RU and EN-JA. Interestingly, the "buggy" prompt actually did slightly better!

Tempting as it was to call it a feature not a bug, our safety check leads us to categorize it as insignificant and discardable statistical noise. We add this to the growing list of observations of rather unpredictable sensitivity of LLMs' output to the fine details of the prompts in some cases, and their surprising robustness to prompt changes in other cases. We further hypothesize that LLMs' insensitivity to the potentially misleading second occurrence of 'Russian' (or 'Japanese') in the "buggy" prompt may have to do with (i) their default preference for English; and/or (ii) their ability to identify the language of the string that actually follows ':'; and/or (iii) the fact that transformer-based neural networks, unlike the older LSTM- and GRU-based architectures, compute the attention scores between all pairs of tokens in the entire input directly and in parallel, rather than consecutively, so the fact that the second occurrence of the language name ('Russian' or 'Japanese') immediately precedes the source sentence does not make the former more important than the other preceding tokens.

We release all translation outputs from the systems we tested in the form of a single Excel file named `1-10_en-ru-ja_short_MT-LLM-outputs.xlsx`, where Column A contains the segment IDs, Column F the source segment length (in words), and the other columns are labeled with the target language and the system which generated the output.

4 Automatic Quality Evaluation

In this section we report and discuss the automatic metric scores for the entire corpus.

As already noted in Section 2 above, we had to reduce the length of our corpus by about 9% to 3555 segments to meet the file size requirements of MATEO (Vanroy et al., 2023), the tool we utilized to calculate the BLEU, chrF2, TER, and COMET scores for our outputs. We provide additional details in Appendix D.

The evaluation scores for `1-8_en` are represented in Table 1 and Figure 1 below. Consistently with other reports, the string-based scores for EN-JA are lower than for EN-RU. We note, however, that the COMET scores are neck-to-neck; in fact, slightly higher for EN-JA for all LLM outputs and MT3. All the score differences are statistically significant.

Since the distinction between the linguistic concepts of character and word is blurred in Japanese, questions may be raised about the separate significance of chrF for translation directions involving this language. We do not have a considered view on this. But we calculated pairwise Pearson correlation values for BLEU-chrF2, BLEU-TER, and BLEU-COMET between the scores for our six systems for both language pairs (Table 10). We note high correlations between BLEU and chrF2, and between BLEU and TER for both language pairs, a somewhat lower but still solid correlation between BLUE and COMET for EN-RU, and the lack of correlation between BLUE and COMET for EN-JA. Along with COMET's neck-to-neck results for both language pairs, this underscores the importance of neural-based metrics.

In our experiments, performed on a 13th Gen Intel(R) Core(TM) i9-13900KF 3.00 GHz 64.0 GB PC, MATEO took roughly 20 minutes to compute the four scores for a single output against a reference; for JA it took slightly longer than for RU. Of these 20 minutes, roughly 16 minutes go into computing the COMET scores, 2 minutes into TER, and 2 minutes into bootstrap resampling at the very end. The calculation of BLUE and chrF2 is very fast. In light of the above-noted considerations, the time spent on computing the COMET scores is the time well spent. Users should be aware of this.

5 What Sample Size is Needed for Reliable Automatic Quality Evaluation?

Another important question that may arise for freelancers inclined to use automatic evaluation of MT/LLM outputs in choosing the best system for a new project is the minimal size of a sample required to make a reliable decision. A freelancer may have a good TM from a previous project in the same domain or for the same client that could be used for reference. Alternatively, a freelancer may complete a representative part of a new project and decide to add the best-performing MT and LLM-based system to their workflow going forward. One can imagine similar scenarios. Such deliberations should, of course, take into account typological differences between target languages which may affect the automatic scores for string- and neural-based metrics differently. In all cases of this sort,

the size of the sample to be used for MT/LLM translation quality evaluation must be statistically significant. What is the minimal size that meets this requirement?

To approach this question empirically we generated additional sets of automatic MT/LLM evaluation scores for the outputs from three distinct parts of our corpus, `229_en` (229 segments, identical to `5_en_short`), `1143_en` (1143 segments, identical to `3_en_short`), and `2183_en` (2183 segments comprising the rest of `1-8_en_short`) to see how well they correlate with each other. Tables 11 and 12 in Appendix E feature the four sets of scores, including those for `1-8_en_short` (3555 = 228 + 1143 + 2183 segments).

The lack of overlap among `229_en`, `1143_en`, and `2183_en` (cumulatively comprising the entire `1-8_en_short` document), which is evidenced in our memoQ analysis (Table 3) makes them suitable for correlation analysis, as does their thematic coherence: all three originate in a single narrow-domain document. Tables 13 and 14 (Appendix F) represent the Pearson correlation values r along with their p-values for three pairs of evaluation scores sets corresponding to `229_en`, `1143_en`, and `2183_en`.

We observe that the correlations are very strong in all cases, across all the metrics. We are thus led to conclude that computing automatic scores for a small part of our document (229/3555 = 6.4%) would give us a good sense of the relative performance of several MT/LLM systems. However, this approach has its limitations. See Appendix F where we also provide additional details regarding the use of statistical methods for freelancers and discuss the prospects for future work.

6 Manual Evaluation of Select Translation Outputs

We selected 180 and 360 MT- and LLM-translated sentences from the outputs for `5_en_short` and `3_en_short` respectively for each of our language pairs (i.e. 1080 segments overall) to perform manual evaluation of their quality with a simple linear scale in order to estimate whether the outputs' sentence-level COMET scores correlate with our "human grades" for them. Below we outline our selection process, the evaluation scale, and the results.

6.1 Segment selection

We ranked MT- and LLM-generated translations by their sentence-level COMET scores and selected 10 highest-scoring segments, 10 intermediate-scoring segments, based on their median ranks, and 10 lowest-scoring segments from the outputs for `5_en_short`. We doubled these numbers (20-20-20) for `3_en_short`.

6.2 Human grading

To assign "human grades" to the selected translations we adopted a linear 0.0–4.0 scale modeled after academic grading (Table 15 in Appendix G). Two of the co-authors who have extensive academic teaching experience found this approach intuitive and efficient: it is easy for them to imagine they are grading student work. Along with the letter/numeric grades, we supplied brief notes for each graded translation highlighting 1–2 most serious issues from the following list: Accuracy; Clarity; Consistency; Fluency; Grammar (including spelling, typography, and syntax); Register; Style; Terminology; Tone. To minimize our bias in grading, we sorted these segments by their ID numbers rather than by their COMET scores.

While we are fully aware of the multiple limitations of this approach, our primary goal in this first round of baseline evaluations was to develop and offer to fellow translators a potentially fruitful method that would allow them to see whether automatic scores correlate with their human judgment in their particular use case.

6.3 Are automatic evaluation scores correlated with human grades?

We calculated Pearson (r) and Spearman (ρ) correlation coefficients between sentence-level COMET scores and our numeric grades for the 10-10-10 and 20-20-20 selections from each translation output (Tables 16 and 17 in Appendix H) for each language pair. Most of the r and ρ values suggest moderate to strong correlation; but the variation is rather wide, between MT/LLM outputs, language pairs, and sample sizes. Some of the variation may an artifact of our somewhat impressionistic and non-rigorous grading and/or the sampling method. These may be adjusted depending on the available human resources. But calculating sentence-level correlations is a very natural and easy strategy to pursue in all cases where "human grades" of select outputs are available.

	COMET	BLEU	chrF2	TER
MT1	88.1	41.1	64.4	43.1
MT2	**90.8**	**57.2**	**74.2**	**31.1**
MT3	90.2	45.4	67.4	40.0
LLM1	88.8	38.4	63.5	45.4
LLM2	89.3	37.1	63.0	46.2
LLM3	88.6	33.2	60.1	50.1

English-Russian

	COMET	BLEU	chrF2	TER
MT1	88.1	31.1	39.5	55.3
MT2	89.7	**38.6**	**46.0**	**47.5**
MT3	**90.6**	36.8	44.1	49.7
LLM1	89.5	31.9	38.6	53.0
LLM2	90.1	30.2	37.6	53.9
LLM3	89.5	28.9	36.3	55.2

English-Japanese

Table 1: Evaluation metric scores for MT and LLM models for English-Russian and English-Japanese translations for `1-8_en_short`.

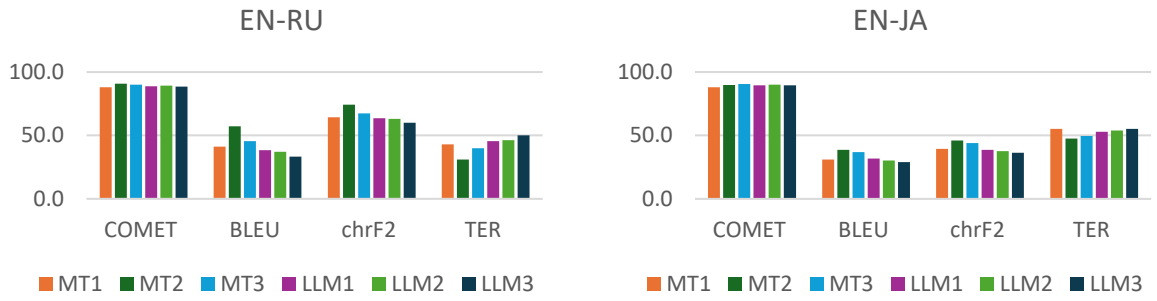


Figure 1: Visualization of MATEO-generated metric scores for EN-RU and EN-JA translations, broken down by MT engine and LLM, for `1-8_en_short`.

We release all the selected sentences along with their COMET scores and our grades and comments in the form of two Excel files: `3_en-ru-ja_short_comet_grades.xlsx` and `5_en-ru-ja_short_comet_grades.xlsx`.

7 Limitations of Our Study and Future Plans

Translation directions. We had an opportunity to experiment with two interestingly different language pairs because we ourselves produced the translations for them and the client gave us permission to use them. We hope that other language directions and document families will be added to our corpus in the future.

Automatic metrics. We limited our choice of them to BLEU, chrF2, TER, and COMET, to maximize efficiency and ease of use. More advanced users should consider other metrics and consult the current best practices (Kocmi et al., 2024).

Correlation experiments with sample sizes reported in Section 5 need to be complemented with power analysis to determine the minimal size of the statistically significant sample. Ideally, future experiments should also include other contrasting pairs—from different domains, registers etc.

Human evaluation requires independent raters and a uniform blinding and randomization protocol. While extensive, our reported results must be taken with a grain of salt. We do believe they serve as a proof of concept.

In future experiments with our corpus (RFTC) we want to explore the potential of various dedicated systems and LLMs for (i) extracting a bilingual glossary from a set of parallel sentences, and (ii) using a glossary thus obtained to improve the quality of translation in the context of the freelancer's workflow.

Other ideas are briefly described in Section 1.1 above. We may pursue some of them and invite fellow translators and other interested parties to join us in this effort.

8 Conclusions

This study demonstrates the potential of Translation Analytics to help freelance translators and smaller language service providers (LSPs) thrive in a rapidly evolving industry. By adapting evaluation

metrics such as BLEU, chrF, TER, and COMET to individual workflows, we provide methods for assessing MT and LLM outputs with rigor and precision. The findings underscore several critical insights:

Utility of automatic evaluation metrics. Automatic metrics, particularly COMET, consistently align with human assessments, reinforcing their value as robust tools for translation quality evaluation. Translators can confidently leverage these metrics to make informed decisions about incorporating MT and LLM systems into their workflows.

Efficiency of sample-based evaluation. Even small, strategically selected samples of documents can yield statistically reliable insights into the relative performance of different translation systems. This approach enables resource-efficient evaluation for freelancers working on large-scale projects.

Integration of human judgment. While automatic metrics are helpful, the integration of human evaluation, anchored in linguistic expertise, remains critical. Our experiments validate the complementary roles of human judgment and automated tools in achieving nuanced and accurate quality assessments.

Empowering freelancers. By demystifying technical methods and tools, we equip translators with the confidence and skills to engage proactively with advanced language technologies. We hope this will help them move beyond being mere participants in the workflow to assuming leadership in optimizing and innovating translation practices. We offer one concrete entry point, with examples of expanded capabilities, in Appendix I.

Future work will focus on expanding the corpus to include additional language pairs, domains, and registers to further validate and refine our methods. Moreover, exploring advanced techniques such as glossary extraction, domain-specific adaptation, reference-free quality estimation, automatic post-editing, and more sophisticated multi-step operations using LLMs represents promising avenues for enhancing translation quality and efficiency.

As the landscape of translation continues to evolve, it is imperative for freelance translators and smaller LSPs to embrace new tools and methodologies. By doing so, they can not only adapt to the changes but also seize the opportunities presented by advancements in language technology. This proactive approach will ensure that translators remain at the forefront of a profession that is as dynamic as it is indispensable.

Author Contributions

YB developed the initial plan, prepared the data, ran the evaluation experiments, and wrote most of the content including literature review and bibliography, but excluding Sections 3.2–3.5 and Appendix I, which were contributed by AB, who took care of all our programming needs. YB and and SFK are ATA-certified translators who worked with their partners in summer 2024 on translating the International Edition of the Reeve Foundation's Paralysis Resource Guide to Russian and Japanese. They performed manual evaluation of the 1080 selected MT- and LLM-generated segment translations as described in Section 6. They also provided additional notes on the RFTC corpus in Appendix B. SFK curated the Japanese portion of the data.

Sustainability Statement

Our experiments performed on personal computers did not involve training of neural models. Computing COMET scores and querying LLMs for translation were the longest operations. We report the runtimes for them in Section 4 and Table 5.

We used one of the recommended algorithms[5] to estimate a carbon impact of our computations according to (Lannelongue et al., 2021). A brief report is included it Appendix J.

Acknowledgments

YB's work is supported by the NSF grant No. SES-233671. We are grateful to Bran Vanroy for helpful comments and clarifications on MATEO. We thank the reviewers for their very helpful comments. We reiterate our thanks to the Christopher & Dana Reeve Foundation for the permission to use their linguistic resources in our experiments.

References

Mohammed Al-Batineh and Moza Al Tenaijy. 2024. Adapting to technological change: An investigation of translator training and the translation market in the Arab world. *Heliyon*, 10(7). Publisher: Elsevier.

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.

[5] https://calculator.green-algorithms.org/

Nathaniel Berger, Stefan Riezler, Miriam Exel, and Matthias Huck. 2024. Prompting large language models with human error markings for self-correcting machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 636–646, Sheffield, UK. European Association for Machine Translation (EAMT).

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.

Bureau Works. 2025. Augmented translation actions: Enhancing translation efficiency with ai. Accessed: 2025-01-20.

Sheila Castilho, Clodagh Quinn Mallon, Rahel Meister, and Shengya Yue. 2023. Do online machine translation systems care for context? what about a GPT model? In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 393–417, Tampere, Finland. European Association for Machine Translation.

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).

Adria Crangasu. 2025. How is Artificial Intelligence Changing the Translation Services Industry?

Agustín Da Fieno Delucchi, Alfredo de Almeida, and Jorge Russo dos Santos. 2025. How Language Industry Jobs Are "Shifting Left". *Multilingual*, (January).

Qiuyu Ding, Hailong Cao, Zihao Feng, Muyun Yang, and Tiejun Zhao. 2025. Enhancing bilingual lexicon induction via harnessing polysemous words. *Neurocomputing*, 611:128682.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Tear: Improving llm-based machine translation with systematic self-refinement. *Preprint*, arXiv:2402.16379.

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. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. *Preprint*, arXiv:2302.01398.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *Preprint*, arXiv:2302.07856.

Jack Halpern. 2025. Language resource action guide. Accessed: 2025-01-20.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

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.

Intento, Inc. 2024. Machine translation report 2024. Accessed: 2025-01-20.

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. *Preprint*, arXiv:2301.08745.

Dayeon Ki and Marine Carpuat. 2024. Guiding large language models to post-edit machine translation with error annotations. *arXiv preprint arXiv:2404.07851*.

Rebecca Knowles and Chi-kiu Lo. 2024. Calibration and context in human evaluation of machine translation. *Natural Language Processing*, pages 1–25.

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 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.

Muhammed Yusuf Kocyigit, Eleftheria Briakou, Daniel Deutsch, Jiaming Luo, Colin Cherry, and Markus Freitag. 2025. Overestimation in LLM Evaluation: A Controlled Large-Scale Study on Data Contamination's Impact on Machine Translation. _eprint: 2501.18771.

Séamus Lankford, Haithem Afli, and Andy Way. 2023. adaptNMT: an open-source, language-agnostic development environment for neural machine translation. *Language Resources and Evaluation*, 57(4):1671–1696.

Loïc Lannelongue, Jason Grealey, and Michael Inouye. 2021. Green Algorithms: Quantifying the Carbon Footprint of Computation. *Advanced Science*, 8(12):2100707.

Chen Li, Meishan Zhang, Xuebo Liu, Zhaocong Li, Derek Wong, and Min Zhang. 2024a. 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.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024b. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

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. 2024. 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, 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.

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.

memoQ. 2025. memoq agt: Advanced machine translation technology. Accessed: 2025-01-20.

Joss Moorkens, Andy Way, and Séamus Lankford. 2025. *Automating translation*. Routledge introductions to translation and interpreting. Routldege, Abingdon, Oxon ; New York, NY.

Yasmin Moslem. 2024. Language modelling approaches to adaptive machine translation. *Preprint*, arXiv:2401.14559.

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.

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.

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. *Preprint*, arXiv:2303.13780.

Ben Peters and André F. T. Martins. 2024. Did translation models get more robust without anyone even noticing? *Preprint*, arXiv:2403.03923.

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.

Celia Rico Pérez. 2024. Re-thinking machine translation post-editing guidelines. *The Journal of Specialised Translation*, 41:10–29.

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, 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)*, pages 2685–2702, Online. Association for Computational Linguistics.

Miguel Rios. 2024. Instruction-tuned large language models for machine translation in the medical domain. *Preprint*, arXiv:2408.16440.

RWS Group. 2025. Beyond words: Exploring the future of translation and localization. Accessed: 2025-01-20.

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.

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.

Slator. 2024. How AI Impacts Jobs, Skills, and Tools for Localization Professionals. Accessed: 2025-04-03.

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.

Bram Vanroy, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine translation evaluation online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500, Tampere, Finland. European Association for Machine Translation.

Inacio Vieira, Will Allred, Séamus Lankford, Sheila Castilho, and Andy Way. 2024. How much data is enough data? fine-tuning large language models for in-house translation: Performance evaluation across multiple dataset sizes. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 236–249, Chicago, USA. Association for Machine Translation in the Americas.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Narcisse Zekpa and Ajeeb Peter. 2025. Evaluate large language models for your machine translation tasks on AWS. Accessed: 2025-01-20.

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*.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

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. *Preprint*, arXiv:2402.15061.

Haotian Zhu, Denise Mak, Jesse Gioannini, and Fei Xia. 2020. NLPStatTest: A toolkit for comparing NLP system performance. 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: System Demonstrations*, pages 40–46, Suzhou, China. 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.

A NMT Engines and LLMs Used in our Experiments

MT1 = ModernMT Professional
`https://www.modernmt.com/translate`
MT2 = DeepL Translator Pro
`https://www.deepl.com/en/translator`
MT3 = Google MT (Cloud Basic)

https://translate.google.com
LLM1 = GPT-4o
https://platform.openai.com/docs/models/gpt-4o
LLM2 = Claude 3.5 Sonnet
https://console.anthropic.com
LLM3 = Gemini 1.5 Pro
https://ai.google.dev/gemini-ap

We selected these engines and models at the time of conducting our baseline experiments (November 2024 – January 2025) based on a balance of the following considerations:

- their popularity among freelance translators and LSPs with limited resources;
- their subscription and per-token costs;
- their existing integration into CAT tools.

There are numerous other options available, including new NMT systems and the latest LLMs, and we plan to explore some of them in the future. We also believe that at the time of our initial experiments reported here, popular LLMs and NMT systems have not seen our trilingual data and, hence, could not have used it for re-training or fine-tuning. Now that this data is available, it might be of some interest to see if our chosen models' performance has changed (Kocyigit et al., 2025).

B The Reeve Foundation Trilingual Corpus: Additional Details

The source document statistics for our corpus are compiled in Table 2 below. Table 3 presents memoQ analyses of both inputs, "long" and "short." Table 4 provides further details of the corpus.

Although translations of the Reeve Foundation International Edition of the Paralysis Resource Guide (PRG) are intended to be generally available, their main target is the US population for whom English is a second language.

In the Russian translation of PRG, organization and program names and most of their acronyms were translated on their first occurrence followed by the English original and acronym in parentheses. In subsequent occurrences in the same section of the document, only translations or translated acronyms (where available) were used. Exceptions include acronyms such as 'FDC' and brand names of companies and their products, such as 'Pfizer' and 'Tobii Dynavox', which are kept in English. The brand medication names were translated or transliterated followed by their original English names on their first appearance. Only translations were used on subsequence occurrences. Number notation generally follows Russian conventions, i.e. '33,000' → '33 000'; '6.79' → '6,79'; etc.

The Japanese translation of PRG generally adheres to the notation guidelines outlined in JTF Style Guide for Translators Working into Japanese. A polite and neutral style using the *desu/masu* form was applied, and the honorific suffix *san* was added after the names of individuals outside the Reeve Foundation. All personal names were transliterated. For medical terms, the original English term and its Japanese translation were juxtaposed in the headings of each section, separated by a slash, while only the Japanese versions were used in the body of the text. In the resource sections, organization names are presented in Japanese first, followed by the original English in parentheses. In the main body text, however, they are only in Japanese. When the source text includes abbreviations or acronyms that may be unfamiliar to Japanese readers, the full form is translated into Japanese. Physical and email addresses, URLs, and phone numbers are left in their original English form.

For typographic conventions, half-width characters are used for Arabic numerals, the percentage sign, slashes for fractions and acronyms, and colons (where unavoidable). Full-width characters are used for exclamation marks, question marks, Japanese middle dots, slashes (except in the cases mentioned above), ampersands, and parentheses. The UTF-8 encoding used in our experiments preserves all the relevant features of Japanese grammar and notation.

C Runtime Details

C.1 MT and LLM translation runtimes and costs

The available runtime and cost details for our translation operations are provided in Table 5 below.

C.2 "Bug-free" vs. "buggy" prompts

As noted in Section 3.6, we generated two sets of LLM outputs for `5_en_short` (229 segments) with the "bug-free" prompt to compare them with the outputs for the "buggy" prompt and computed their automatic scores with MATEO. See Tables 6 and 7 below.

D Automatic Quality Evaluation Details

As noted in Section 1.3, evaluating the quality of MT output is a central concern in research and

development. Automatic MT quality metrics are tools that help measure how good an MT-translated sentence is, typically by comparing it to one or more human reference translations. While the field is rapidly evolving, several widely used metrics include BLEU, chrF, TER, and COMET. These can be broadly categorized into *string-based* and *neural-based* metrics.

String-based metrics evaluate translations by comparing the surface forms—words or characters—of MT output and reference human translations. BLEU (BiLingual Evaluation Understudy, Papineni et al., 2002) is one of the earliest and most well-known metrics. BLEU calculates how many n-grams (word sequences) in the MT output match those in the reference. While useful, it can be overly strict, penalizing valid translations that use synonyms or different phrasing. chrF (Character F-score, Popović, 2015), on the other hand, operates at the character level, making it more sensitive to morphologically rich languages and spelling. It computes F-scores based on overlapping character n-grams, which helps in capturing partial matches more effectively. TER (Translation Edit Rate, Snover et al., 2006) metric measures the number of edits (insertions, deletions, substitutions, and shifts) needed to change the MT output into the reference translation. A lower TER indicates better translation quality. It gives a more intuitive sense of the editing effort required.

Neural-based metrics leverage LLMs and machine learning techniques to evaluate translations more like humans do. These models can understand meaning beyond surface similarity. One such metric, used in this paper, is COMET (Crosslingual Optimized Metric for Evaluation of Translation, Rei et al., 2020). Built on pre-trained neural models and fine-tuned on human quality assessments, it can capture semantic similarity and fluency better than traditional metrics, even when there is little word overlap. It has been shown to correlate better with human judgments. Our results reported in Section 6.3 are consistent with this claim.

Machine translation evaluation is a fast-moving area of research, with new methods and tools emerging regularly. A great place to stay updated is `http://www.machinetranslate.org`, which offers accessible summaries of research, tools, and best practices in the field.

As noted in Section 1.3, one exciting development for practitioners is that freelance translators and non-specialists can now use web-based tools to evaluate MT output themselves. MATEO (Vanroy et al., 2023), employed in our work, is a user-friendly Streamlit-based platform that allows anyone to calculate multiple MT quality metrics—including BLEU, chrF, TER, and COMET—without needing technical knowledge.

There is thus no mystery to MT quality evaluation. These metrics, whether simple or complex, are just tools to help us understand how well an NMT engine or an LLM has translated a piece of text. As the tools become more accessible and sophisticated, translators and content creators are empowered to make informed decisions about using and improving MT output.

A point of caution: when utilizing MATEO, or any other toolkit, for MT evaluation, it is crucial to select the appropriate *metric configurations* to ensure accurate and meaningful results. While the default settings in MATEO are designed to be practical for a wide range of target languages, evaluating translations into morphologically rich languages, such as Japanese or Korean, requires special attention. These languages exhibit complex word forms and inflections that standard metric configurations might not fully capture.

MATEO allows one to make the necessary changes in the "Metric selection" section of the application. In our case, evaluation of the outputs in Japanese required changing the default tokenization setting in BLEU to `ja-mecab`, and enabling `asian-support` in TER. We also found it useful to enable the `normalized` mode in TER, which is set to 'False' by default.

We report the configurations we used for the automatic evaluation metrics for both our target languages in Tables 8 and 9.

E Evaluation Scores for Sub-Documents of Different Sizes

See Tables 11 and 12 below, which represent the four sets of automatic metric scores for three non-overlapping parts of `1-8 en short` along with the whole: 3555 = 228 + 1143 + 2183 segments.

F Pearson Correlations for Three Pairs of Score Value Sets Across Translation Systems

Pearson's correlation coefficient (r) is a statistical measure that quantifies the strength and direction of

the linear relationship between two variables, helping to determine whether changes in one variable are associated with changes in another.

The formula for calculating Pearson's correlation is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where x_i and y_i are individual data points, $\bar{x}$ and $\bar{y}$ are their means, and n is the number of data pairs. In essence, r measures how much two variables change together relative to how much they change individually. The numerator represents the covariance between the variables, while the denominator normalizes this value using the standard deviations of both variables.

The value of r always lies between -1.0 (corresponding to perfect negative correlation) and 1.0 (perfect positive correlation). Typical guidelines for interpreting the values of r are as follows:

r	**Correlation**
0.9 to 1.0 or -0.9 to -1.0	Very strong
0.7 to 0.9 or -0.7 to -0.9	Strong
0.5 to 0.7 or -0.5 to -0.7	Moderate
0.3 to 0.5 or -0.3 to -0.5	Weak
0.0 to 0.3 or 0.0 to -0.3	Very week or none

The p-value associated with Pearson's correlation r estimates the statistical significance of the observed correlation. A p-value is the probability that the actual distribution of the data points would occur by random chance. A low p-value (typically < 0.05) suggests that the result is statistically significant.

In our case, the variables in questions are pairwise metric-specific scores which are reflected in the rows of Tables 11 and 12. For example, the first two rows in Table 11 (for EN-RU) show the COMET scores for `229_en` and `1143_en` across our six MT/LLM systems. Accordingly: $x_1 = 87.7, x_2 = 91.1, x_3 = 90.0, x_4 = 88.7, x_5 = 89.6, x_6 = 88.9, y_1 = 89.4, y_2 = 91.4, y_3 = 90.8, y_4 = 89.7, y_5 = 90.1, y_6 = 89.1$, yielding $r = 0.891; p = 0.0077$, reflected in the last column of Table 13.

Thus Tables 13 and 14 below display the correlation coefficients and their p-values for three pairs of score value sets for the outputs from our range of six translation systems (i.e. MT1–MT3 and LLM1–LLM3), in both language directions.

The plots in Figure 2 provide the additional details of the distribution of our "data points" across MT1–3 and LLM 1–3.

As we noted in Section 5, all the pairwise Pearson correlations for our three non-overlapping subdocuments are very strong and statistically significant thus highlighting the consistency and stability of the rankings of our MT/LLM outputs across sub-documents of different sizes. If we wanted to select one or two best performing systems based on the automatic evaluation scores for our project, we could simply pick out the shortest chapter of PRG (i.e. Chapter 5 = `229_en`) and treat it as a good representative of the entire document.

Even this shortest sample has over 4,000 source words, which exceeds the average daily output of a typical translator. It would be interesting to trim down the sample size even more to determine the point at which the correlation is lost and the scores become unreliable. The best way to do this is to perform a power analysis using one of the available toolkits (e.g. Zhu et al., 2020). It would also be desirable to include contrasting pairs of data points from different translation domains and registers. We leave it for further work.

Translators interested in implementing correlation or more advanced statistical analyses can use any number of generally available tools, from Excel to Python or R libraries. In our experience, LLMs can generate simple standalone Python scripts for such purposes, in response to sufficiently detailed prompts.

G Manual Grading Scale

Our manual scale modeled after academic grading is displayed in Table 15.

H Correlation Between Sentence-Level COMET Scores and Numeric Human Grades

Tables 16 and 17 represent Pearson and Spearman correlation between sentence-level COMET scores and our numeric human grades.

As noted above (Appendix F), Pearson correlation measures the strength and direction of a *linear* relationship between two variables. It assumes that both variables are normally distributed, and that the relationship is linear. This approximation

was adequate for six pairs of data points representing the scores for the outputs of our MT/LLM systems. But the number of our chosen sentence-level COMET scores and the corresponding human grades is larger: 30 or 60. In such cases Pearson correlation may be insufficient, especially if the relationship between variables is non-linear or if the data contains outliers. In such cases, adding Spearman correlation (ρ) can provide a more accurate picture of the association by focusing on the rank-order rather than precise values. Spearman correlation is a non-parametric measure that assesses how well the relationship between two variables can be described by a *monotonic* function. It uses the ranked values of the data, not the raw values, so it doesn't assume normality or linearity.

I APIs and Applied Technical Avenues for Freelancers

We acknowledge that most freelance translators are not programmers. However, as discussed elsewhere in this paper, we believe the future of translation work demands skills that are more conducive to the building blocks of machine intelligence and automation.

As a practical matter, the major LLM providers expose use of their models in two ways: a human-friendly way, via an interactive "chatbot" interface, and a machine-friendly way, via REST (REpresentational State Transfer) APIs, or Application Programming Interfaces. Despite the imposing weight of these acronyms to non-technical readers, the chasm between these modes of interaction is not, in fact, so vast. REST APIs use HTTP, the building-block protocol of the World Wide Web, as a transport, and a series of HTTP chatbot "verbs" whose meaning is not especially obscure: `GET`, `POST`, `DELETE`, and so on.

Contemporary REST APIs customarily encode information in a lightweight, human-readable encapsulation structure known as JSON (JavaScript Object Notation). The primary purpose of JSON is to define a hierarchical relational structure—for instance, to distinguish an object from its attributes.

A simple JSON structure might look like this:

```
{
    "people": [
        "Alex": {
            "org": "Evariste␣Systems",
            "phd": false
        },
        "Yuri": {
            "org": "University␣of␣Georgia",
            "phd": true
        }
    ],
    "paper_type": "edifying",
    "lucky_numbers": [1, 7, 10]
}
```

The SDKs (Software Development Kits) of major LLM providers abstract away lower-level programmatic REST API interactions, which is more ergonomic for the software engineers using them. However, the LLM APIs can be directly queried, with the help of user-friendly tools such as Postman. The interactive "chatbot" clients to which many readers will be well-accustomed are little more than simplified front-ends to these REST APIs.

Perusing the content of JSON responses from the major LLM providers' REST APIs can open one's mind to new possibilities. For example, one of the authors used Postman to prompt OpenAI's GPT-4o model thus:

"You are a highly competent Russian to English translator. How would you explain the Russian concept of a 'matryoshka' in English? Please be brief."

On the surface, the reply received was unremarkable:

```
 {
"id":
    "chatcmpl-BLGsJZWEEWEfIBCkS46cXpNyyTfnY",
"object": "chat.completion",
"created": 1744409455,
"model": "gpt-4o-2024-08-06",
"choices": [
 {
   "index": 0,
   "message": {
     "role": "assistant",
     "content": "A␣matryoshka,␣also␣known␣as␣
          a␣Russian␣nesting␣doll,␣is␣a␣set␣of␣
          wooden␣dolls␣of␣decreasing␣size␣
          placed␣one␣inside␣another.␣Each␣doll␣
          splits␣in␣half␣at␣the␣middle␣to␣
          reveal␣a␣smaller␣doll␣inside,␣
          symbolizing␣themes␣of␣motherhood,␣
          family,␣and␣continuity.",
     "refusal": null,
     "annotations": []
   },
   [...]
```

However, after perusing OpenAI chat API reference, the author learned that it is possible to supply the JSON attributes:

```
"logprobs": true,
"top_logprobs:␣3
```

to the request, which tells OpenAI to share two other alternative probabilistic paths not taken for every generated token.

Thus, although GPT-4o began this generated response with the article 'A', it considered alternatives:


```
"logprobs": {
    "content": [
     {
       "token": "A",
       "logprob": -0.011159946210682392,
       "bytes": [
         65
       ],
       "top_logprobs": [
         {
           "token": "A",
           "logprob": -0.011159946210682392,
           "bytes": [
             65
           ]
         },
         {
           "token": "The",
           "logprob": -4.511159896850586,
           "bytes": [
             84,
             104,
             101
           ]
         },
         {
           "token": "In",
           "logprob": -9.636159896850586,
           "bytes": [
             73,
             110
           ]
         }
       ]
     },
```

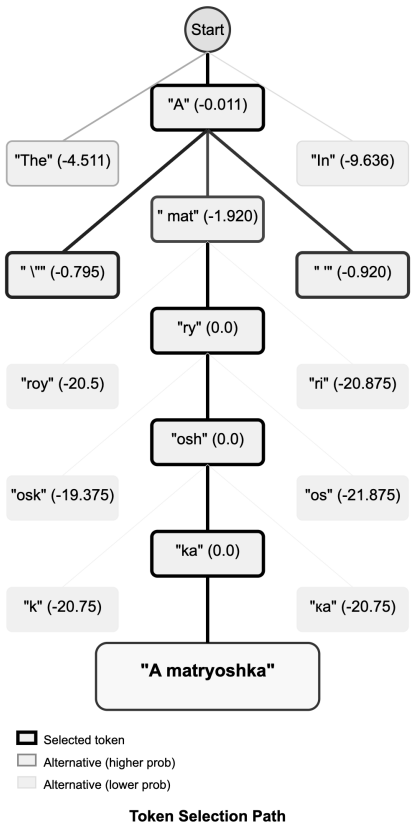

The author fed this probability output for the first few tokens into Anthropic's Claude Sonnet model and asked it to generate a flowchart, using the following prompt:

> "This JSON file contains a 'logprobs' element from OpenAI's API, which shows two other probabilistic responses considered by the model before returning the one with the lowest absolute value. Could you please draw a flowchart for the first five 'logprobs' elements which illustrates the traversal path taken? Please encapsulate every one of the five generated tokens in a rectangle, and use a darker or solid line to indicate the path actually taken based on the lowest absolute value of the 'logprob' entry, while using lighter lines, their lightness in proportion to the relative absolute value of the 'logprob' value, to show alternatives not taken."

The author then further prompted Claude to flatten the graphic for ease of inclusion here:

> "Could you refine this flowchart to be more vertical, so that it is easier to incorporate into a two-column document without overflow beyond the margins?"

This was the result:



A key idea here is that this interesting foray would have never occurred to the author without digging into the OpenAI APIs and interacting with them directly. The commonplace interactive chatbot interfaces do not surface these possibilities to the end-user. No code was written for this exercise, just some slight tweaks to minimal, easy-to-read JSON data structures.

First and foremost, we believe that becoming conversant with the API surface of the major LLM providers can empower freelancers to make more specific and technically articulate LLM integration demands of the vendors of their preferred translation software tools. Second, we can reasonably speculate that the creation or enhancement of tools

reliant on API integrations may drift further away from the exclusive province of professional programmers, and become more reachable for technically minded end-users. This trend can be extrapolated from an ongoing trajectory to which veteran software engineers are privy: service APIs offered over the Internet have become far less arcane and easier to decipher over time, between the simplified vocabulary of REST and the human-readable wire format of JSON, for example.

J Carbon Imprint

We used the Green Algorithm developed in (Lannelongue et al., 2021) to estimate the carbon imprint of our computations performed on two computers (Figure 3 and Figure 4 below).

	1-10_en_long	1-10_en_short	1-8_en_short
Segments	**4528**	**3896**	**3555**
Words tokens (no punc)	76,553	74,667	68,989
Word types (no punc)	10,689	10,325	9,821
Characters (w/o \r\n)	500,347	485,830	448,652
Type/token ratio	0.14	0.14	0.14
MTLD	100.18	101.44	100.77
Average segment length (words)	16.91	19.17	19.41
Average word length (characters)	5.42	5.38	5.38

Table 2: Source document statistics.

Type	Segments.	Source words	Source chars	Source tags	Percent
		1-10_en_short			
All	**3896**	74683	415016	0	100
X-translated / double context	0	0	0	0	0
Repetition	0	0	0	0	0
101%	0	0	0	0	0
100%	0	0	0	0	0
95%–99%	3	31	227	0	0.04
85%–94%	6	118	696	0	0.16
75%–84%	17	251	1703	0	0.34
50%–74%	201	3080	18615	0	4.12
No match	3669	71203	393775	0	95.34
		1-8_en_short			
All	**3555**	68999	383178	0	100
X-translated / double context	0	0	0	0	0
Repetition	0	0	0	0	0
101%	0	0	0	0	0
100%	0	0	0	0	0
95%–99%	3	31	227	0	0.04
85%–94%	6	118	696	0	0.17
75%–84%	16	239	1636	0	0.35
50%–74%	172	2700	16465	0	3.91
No match	3358	65911	364154	0	95.52

Table 3: MemoQ analysis of the "long" and "short" PRG inputs. Fuzzy matches result from the *Homogeneity* feature of the memoQ analysis which measures internal similarities within a set of documents by adding each segment to a temporary TM and using it for lookup for every subsequently processed segment. For details, see here.

Document Name	Excel/PDF	Notes
PARALYSIS RESOURCE GUIDE (PRG): International Edition	1-10_en-ru-ja_long.xlsx 1-10_en-ru-ja_short.xlsx PRG-IntEd_en.pdf PRG-IntEd_ru.pdf PRG-IntEd_ja.pdf	4528 segments (ID: 3–4687) 3896 segments (ID: 13–4687)
PRG: Chapters 1-6	1-8_en-ru-ja_short.xlsx	3555 segments (ID: 13–4299)
PRG: Front Cover		No segments
PRG: Introduction	2_en-ru-ja_short.xlsx	21 segments (ID: 13–45)
PRG: Chapter 1	3_en-ru-ja_short.xlsx	1143 segments (ID: 108–1430)
PRG: Chapter 2	4_en-ru-ja_short.xlsx	1237 segments (ID: 1437–2897)
PRG: Chapter 3	5_en-ru-ja_short.xlsx	229 segments (ID: 2903–3178)
PRG: Chapter 4	6_en-ru-ja_short.xlsx	228 segments (ID: 3186–3467)
PRG: Chapter 5	7_en-ru-ja_short.xlsx	583 segments (ID: 3473–4156)
PRG: Chapter 6	8_en-ru-ja_short.xlsx	114 segments (ID: 4164–4299)
PRG: Glossary	9_en-ru-ja_short.xlsx	335 segments (ID: 4306–4673)
PRG: Back Cover	10_en-ru-ja_short.xlsx	6 segments (ID: 4675–4687)

Table 4: Document details for the Paralysis Resource Guide (PRG).

		Runtime	Total cost	Notes
MT1	EN-RU	00:02:17	n/a	
	EN-JA	00:02:32	n/a	
MT2	EN-RU	00:01:28	n/a	
	EN-JA	00:01:05	n/a	
MT3	EN-RU	00:08:36	n/a	
	EN-JA	00:08:03	n/a	
LLM1	EN-RU	00:15:38	USD 7.01	Combined EN-RU and EN-JA
	EN-JA	00:15:28		
LLM2	EN-RU	02:29:43	USD 23.22	Combined EN-RU and EN-JA
	EN-JA	02:38:38		
LLM3	EN-RU	00:11:25	n/a	
	EN-JA	00:11:25		

Table 5: MT and LLM translation runtimes and costs for `1-10_en_short` (3896 segments).

	COMET	BLEU	chrF2	TER
"Bug Free-1"	89.1 ± 0.4	33.6 ± 1.7	60.7 ± 1.2	49.6 ± 1.5
"Bug Free-2"	89.0 ± 0.5	33.7 ± 1.7 ($p = 0.23$)*	60.7 ± 1.1 ($p = 0.33$)*	49.6 ± 1.5 ($p = 0.33$)
"Buggy"	89.1 ± 0.5 ($p = 0.42$)	34.1 ± 1.7 ($p = 0.08$)	61.0 ± 1.2 ($p = 0.04$)*	49.3 ± 1.5 ($p = 0.09$)

Table 6: LLM1–3 combined outputs for `5_en_short`: English-Russian.

	COMET	BLEU	chrF2	TER
"Bug Free-1"	89.7 ± 0.4	29.8 ± 1.3	36.8 ± 1.3	52.0 ± 1.3
"Bug Free-2"	89.7 ± 0.4 ($p = 0.34$)	29.6 ± 1.3 ($p = 0.15$)	36.4 ± 1.2 ($p = 0.05$)*	52.2 ± 1.3 ($p = 0.09$)
"Buggy"	89.9 ± 0.4 ($p = 0.05$)*	30.5 ± 1.3 ($p = 0.02$)*	37.4 ± 1.3 ($p = 0.04$)*	51.6 ± 1.3 ($p = 0.05$)

Table 7: LLM1–3 combined outputs for `5_en_short`: English-Japanese.

Metric	Details
BLEU	`nrefs:1|bs:1000|seed:12345|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1|mateo:1.1.3`
chrF2	`nrefs:1|bs:1000|seed:12345|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1|mateo:1.1.3`
TER	`nrefs:1|bs:1000|seed:12345|case:lc|tok:tercom|norm:yes|punct:yes|asian:no|version:2.3.1|mateo:1.1.3`
COMET	`nrefs:1|bs:1000|seed:12345|c:Unbabel/wmt22-comet-da|version:2.0.1|mateo:1.1.3`

Table 8: Metrics configurations for English-Russian.

Metric	Details
BLEU	`nrefs:1|bs:1000|seed:12345|case:mixed|eff:no|tok:ja-mecab-0.996-IPA|smooth:exp|version:2.3.1|mateo:1.1.3`
chrF2	`nrefs:1|bs:1000|seed:12345|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1|mateo:1.1.3`
TER	`nrefs:1|bs:1000|seed:12345|case:lc|tok:tercom|norm:yes|punct:yes|asian:yes|version:2.3.1|mateo:1.1.3`
COMET	`nrefs:1|bs:1000|seed:12345|c:Unbabel/wmt22-comet-da|version:2.0.1|mateo:1.1.3`

Table 9: Metrics configurations for English-Japanese.

		BLEU-chrF2	BLEU-TER	BLEU-COMET
EN-RU	r	0.998	-0.999	0.806
	p	<0.0001	<0.0001	0.0345
EN-JA	r	0.990	-0.969	0.388
	p	<0.0001	0.0002	0.4321

Table 10: Pearson correlations (r) for BLEU-chrF2, BLEU-TER, and BLEU-COMET for 1-8_en.

Metric	Label	Sgmts	MT1	MT2	MT3	LLM1	LLM2	LLM3
COMET	229_en	229	87.7	91.0	90.0	88.7	89.6	88.9
	1143_en	1143	89.4	91.4	90.8	89.7	90.1	89.1
	2183_en	2183	87.5	90.5	89.9	88.4	88.9	88.3
	1-8_en	3555	88.1	90.8	90.2	88.8	89.3	88.6
BLEU	229_en	229	37.2	57.7	43.6	34.6	35.3	32.3
	1143_en	1143	45.8	60.1	49.0	42.5	40.8	36.2
	2183_en	2183	39.0	55.5	43.6	36.5	35.3	31.8
	1-8_en	3555	41.1	57.2	45.4	38.4	37.1	33.2
chrF2	229_en	229	62.2	74.7	65.6	61.5	62.3	59.4
	1143_en	1143	68.3	76.4	70.4	67.0	66.2	62.9
	2183_en	2183	62.5	73.0	66.0	61.7	61.4	58.6
	1-8_en	3555	64.4	74.2	67.4	63.5	63.0	60.1
TER	229_en	229	46.0	30.9	42.3	48.7	47.4	51.6
	1143_en	1143	38.7	28.5	36.5	41.4	42.7	47.3
	2183_en	2183	45.2	32.6	41.6	47.2	47.9	51.4
	1-8_en	3555	43.1	31.1	40.0	45.4	46.2	50.1

Table 11: Evaluation scores for documents of different sizes: English-Russian.

Metric	Label	Sgmts	MT1	MT2	MT3	LLM1	LLM2	LLM3
COMET	229_en	229	88.0	90.8	90.3	89.6	90.1	89.8
	1143_en	1143	88.6	89.8	90.8	89.7	90.3	89.8
	2183_en	2183	87.8	89.6	90.5	89.4	90.0	89.4
	1-8_en	3555	88.1	89.7	90.6	89.5	90.1	89.5
BLEU	229_en	229	30.8	36.3	35.7	31.3	30.3	29.8
	1143_en	1143	31.0	35.1	36.7	32.0	29.5	28.9
	2183_en	2183	31.2	40.3	36.9	31.9	30.6	28.9
	1-8_en	3555	31.1	38.6	36.8	31.9	30.2	28.9
chrF2	229_en	229	38.7	43.5	42.7	37.7	37.6	36.7
	1143_en	1143	38.7	42.1	43.4	37.6	36.5	35.4
	2183_en	2183	40.0	47.8	44.6	39.3	38.2	36.8
	1-8_en	3555	39.5	46.0	44.1	38.6	37.6	36.3
TER	229_en	229	54.1	49.2	48.2	50.9	51.4	52.7
	1143_en	1143	53.4	48.8	48.3	51.9	52.6	53.5
	2183_en	2183	56.5	47.0	50.7	53.9	54.9	56.5
	1-8_en	3555	55.3	47.5	49.7	53.0	53.9	55.2

Table 12: Evaluation scores for documents of different sizes: English-Japanese.

EN-RU	Correlation pairs		
	229 / 2183	**1143 / 2183**	**229 / 1143**
COMET			
r	0.978	0.942	0.891
p-value	0.0001	0.0014	0.0077
BLEU			
r	0.991	0.994	0.973
p-value	< 0.0001	< 0.0001	0.0001
chrF2			
r	0.990	0.990	0.966
p-value	< 0.0001	< 0.0001	0.0003
TER			
r	0.992	0.990	0.971
p-value	< 0.0001	< 0.0001	0.0002

Table 13: Pearson correlations (r) and p-values for EN-RU for three pairs of score value sets across six translation systems.

EN-JA	Correlation pairs		
	229 / 2183	**1143 / 2183**	**229 / 1143**
COMET			
r	0.869	0.989	0.797
p-value	0.0126	< 0.0001	0.0387
BLEU			
r	0.980	0.901	0.955
p-value	0.0001	0.0060	0.0007
chrF2			
r	0.984	0.929	0.967
p-value	< 0.0001	0.0024	0.0003
TER			
r	0.852	0.922	0.936
p-value	0.0174	0.0031	0.0018

Table 14: Pearson correlations (r) and p-values for EN-JA for three pairs of score value sets across six translation systems.

Letter grade	**Numeric grade**
A	4.00
A-	3.67
A-/B+	3.50
B+	3.33
B	3.00
B-	2.67
B-/C+	2.50
C+	2.33
C	2.00
C-	1.67
C-/D+	1.50
D+	1.33
D	1.00
F	0.00

Table 15: Manual grading scale.

5_en_short: English-Russian				
	Pearson Correlation		**Spearman Correlation**	
	r	p	ρ	p
MT1	0.689	< 0.0001	0.755	< 0.0001
MT2	0.549	0.0016	0.765	< 0.0001
MT3	0.660	0.0001	0.765	< 0.0001
LLM1	0.692	< 0.0001	0.806	< 0.0001
LLM2	0.795	< 0.0001	0.734	< 0.0001
LLM3	0.548	0.0016	0.567	0.0011

5_en_short: English-Japanese				
	Pearson Correlation		**Spearman Correlation**	
	r	p	ρ	p
MT1	0.609	0.0004	0.743	< 0.0001
MT3	0.626	0.0002	0.699	< 0.0001
MT4	0.738	< 0.0001	0.657	0.0001
LLM1	0.827	< 0.0001	0.811	< 0.0001
LLM2	0.767	< 0.0001	0.783	< 0.0001
LLM3	0.839	< 0.0001	0.910	< 0.0001

Table 16: Pearson (r) and Spearman (ρ) correlation coefficients between sentence-level COMET scores and numeric human grades for the select "10-10-10" translation outputs for 5_en_short.

3_en_short: English-Russian				
	Pearson Correlation		**Spearman Correlation**	
	r	p	ρ	p
MT1	0.927	< 0.0001	0.895	< 0.0001
MT2	0.845	< 0.0001	0.779	< 0.0001
MT3	0.878	< 0.0001	0.851	< 0.0001
LLM1	0.761	< 0.0001	0.830	< 0.0001
LLM2	0.697	< 0.0001	0.738	< 0.0001
LLM3	0.663	< 0.0001	0.774	< 0.0001

3_en_short: English-Japanese				
	Pearson Correlation		**Spearman Correlation**	
	r	p	ρ	p
MT1	0.558	< 0.0001	0.672	< 0.0001
MT3	0.831	< 0.0001	0.843	< 0.0001
MT4	0.630	< 0.0001	0.694	< 0.0001
LLM1	0.582	< 0.0001	0.589	< 0.0001
LLM2	0.462	0.0002	0.436	0.0004
LLM3	0.665	< 0.0001	0.646	< 0.0001

Table 17: Pearson (r) and Spearman (ρ) correlation coefficients between sentence-level COMET scores and numeric human grades for the select "20-20-20" translation outputs for 3_en_short.

Figure 2: Automatic evaluation scores across the outputs of six translation systems for three non-overlapping parts of the RFTC corpus: 229, 1143, and 2183 segments.

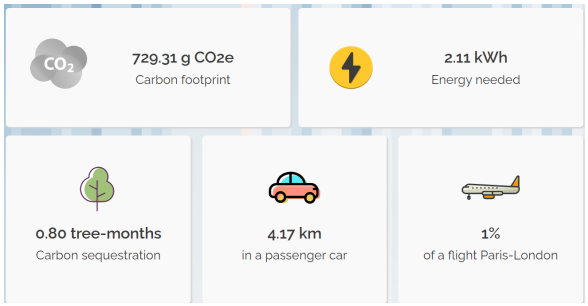


Figure 3: Carbon imprint for 13th Gen Intel(R) Core(TM) i9-13900KF 3.00 GHz 64.0 GB PC.

Figure 4: Carbon imprint for virtual server with Intel Xeon (Skylake) 6-core CPU, 16 GB of RAM.

ITALERT: Assessing the Quality of LLMs and NMT in Translating Italian Emergency Response Text

Maria Carmen Staiano[1], **Lifeng Han**[2,4], **Johanna Monti**[3], **Francesca Chiusaroli**[1]

[1]University of Macerata, [2]LIACS & LUMC, Leiden University
[3]University of Naples "L'Orientale", [4]The University of Manchester
m.staiano@unimc.it, l.han@lumc.nl, jmonti@unior.it, f.chiusaroli@unimc.it

Abstract

This paper presents the outcomes of an initial investigation into the performance of Large Language Models (LLMs) and Neural Machine Translation (NMT) systems in translating high-stakes messages. The research employed a novel bilingual corpus, **ITALERT** (Italian Emergency Response Text) and applied a human-centric post-editing based metric (HOPE) to assess translation quality systematically. The initial dataset contains eleven texts in Italian and their corresponding English translations, both extracted from the national communication campaign website of the Italian Civil Protection Department. The texts deal with **eight crisis scenarios**: *flooding*, *earthquake*, *forest fire*, *volcanic eruption*, *tsunami*, *industrial accident*, *nuclear risk*, and *dam failure*. The dataset has been carefully compiled to ensure usability and clarity for evaluating machine translation (MT) systems in crisis settings. Our findings show that current LLMs and NMT models, such as ChatGPT (OpenAI's GPT-4o model) and Google Translate, **face limitations** in translating emergency texts, particularly in maintaining the appropriate register, resolving context ambiguities, and managing domain-specific terminology. The ITALERT corpus and evaluations are hosted openly at https://github.com/mcstaiano/ITALERT.

1 Introduction

LLMs have shown remarkable advancements in generating fluent and coherent translations. They are trained on large-scale multilingual datasets and can improve translation quality, efficiency and domain adaptation. The interest in LLMs also stems from the fact that they can provide valid translations for high-resource languages, producing competitive results with respect to traditional MT systems (Jiao et al., 2023).

Given their potential, this preliminary evaluation aims to assess the effectiveness of state-of-the-art language models during the preparedness and response phases of an emergency (Lindell et al., 2006) for the language pair **Italian→English** (**RQ**). The rationale for selecting this language pair lies in the need to communicate effectively with the non-Italian-speaking population, including immigrants, refugees, and tourists who are in Italy. The study focuses on identifying which translation system performs better under emergency conditions for specific natural disaster scenarios, providing critical insights to improve multilingual crisis communication in these phases.

The current Italian-English bilingual corpus consists of 13,218 words in total: 6,622 in Italian and 6,596 in English. It includes 440 segments across eight subdomains (flooding, earthquake, forest fire, volcanic eruption, tsunami, industrial accident, nuclear risk, and dam failure), extracted from the communication campaign website *Io non rischio*.[1] Table 1 provides an overview of the corpus, detailing the subdomains along with the word counts for both the original Italian texts and their English translations, sourced directly from the campaign's official website. In addition to the bilingual ITALERT corpus, we present our investigation methodology, and the evaluation of Google MT and ChatGPT using the HOPE metric (Gladkoff and Han, 2022).

2 Background and Related Work

Previous studies of crisis translation have demonstrated that MT systems can be fast, reliable tools for emergency response (Lewis, 2010). The past decade has seen a renewed focus on the possibility of using MT in preventing and mitigating disasters (Federici, 2016; O'Brien and Cadwell, 2017; Federici and Cadwell, 2018). However, few researchers in Italy have addressed the potential of

[1]https://www.iononrischio.gov.it

Proceedings of Machine Translation Summit XX Volume 1, pages 566–577, June 23-27, 2025

Subdomain	IT	EN
Flooding	633	637
Earthquake	368	372
Forest Fire	342	373
Volcanic Eruption	3231	3246
Tsunami	456	445
Industrial Accident	366	366
Nuclear Risk	735	682
Dam Failure	491	475
Total	**6,622**	**6,596**

Table 1: Subdomain-level word counts for the Italian-English bilingual corpus

MT systems in the crisis management workflow. During crises, it is crucial that messages are spread quickly and effectively to the population (Cadwell et al., 2019). But what impact can these messages have if they are communicated in a language that's foreign to the recipients or only partially understood by them?

Based on the latest tourism report provided by ISTAT(ISTAT, 2023), in 2023 Italy recorded 234.2 million overnight stays by foreign tourists, with non-residents making up 52.4% of total hospitality demand. At the same time, the country experienced a significant influx of immigrants and asylum seekers, representing more than 15 nationalities (ISTAT, 2024). In these multilingual and multicultural contexts, English is frequently used as a *lingua franca*, especially in interactions between migrants, institutions, and interpreters (Amato and Cirillo, 2024). This highlights the importance of using English as a vehicular language to foster mutual intelligibility in critical legal, medical, and social settings. For these reasons, we chose the language pair IT→EN for our translations. Our goal is to answer the following **research question**: How accurate are current language models in translating crisis-related texts from Italian to English?

In line with this direction, we started researching existing corpora on the topic of crisis translation in Italy, and we identified one dataset containing humanitarian response documents: HumSet (Fekih et al., 2022). However, this dataset was not selected for our study due to its lack of data in Italian, a critical requirement for our research objectives. While HumSet offers valuable multilingual resources (English, Spanish, French), the absence of Italian significantly limits its relevance to our analysis, which focuses on evaluating the quality of translations in crisis communication scenarios, specifically involving the IT→EN language pair.

Regarding MT evaluations and interpretability, previous work by Han et al. (2021) has examined both human and automatic evaluation methods. Recent research has also explored explainable MT evaluation (Leiter et al., 2024; Perrella et al., 2024), with a focus on providing detailed, interpretable error analyses. Additionally, the human-centric post-editing based metric HOPE offers both explainable feedback and supports the creation of a post-edited gold-standard corpus. As outlined by (Lommel et al., 2024), HOPE adopts a simplified and practical approach to human evaluation, specifically designed for machine translation outputs. Given that our study also aims to develop such a bilingual corpus, HOPE is a suitable choice, as it offers both detailed error analysis and the option to produce a post-edited corpus. The original HOPE metric has eight predefined error categories and severity levels. The eight error types in HOPE are: Impact (IMP), Required Adaptation Missing (RAM), Terminology (TRM), Ungrammatical (UGR), Mistranslation (MIS), Style (STL), Proofreading error (PRF), and Proper Name (PRN). The error severity levels and corresponding point values are: minor (1), medium (2), major (4), severe (8), and critical (16). For our annotation, we decided to use only 7 of the 8 error categories in HOPE, which will be explained in later sections (4.3).

3 Investigation Methodology

The ITALERT methodology for our investigation is shown in Figure 1.

1) The first step consists of data extraction and corpus collection. After the selection of the source texts in Italian, they were segmented into sentences for the MT evaluation phase. The source texts belong to eight subdomains (flooding, earthquake, forest fire, volcanic eruption, tsunami, industrial accident, nuclear risk, and dam failure) extracted from the national communication campaign website of the Italian Civil Protection Department (*Io non rischio*).

2) The second step involves selecting two MT systems and carrying out the automatic translation. Here, we aim to investigate Generative AI models (using ChatGPT-4o) and the standard NMT models (using Google MT) (Johnson et al., 2017). For ChatGPT, we used a zero-shot prompting technique

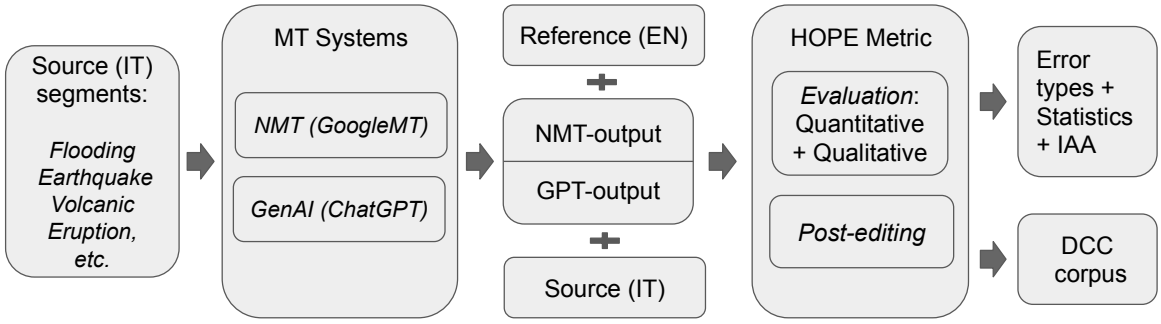


Figure 1: The ITALERT Investigation Methodology Design Framework.

(Cui et al., 2023), asking the model to perform a translation task. The *prompt* was as follows: "This is a document in Italian including crisis or emergency text for eight subdomains, specified in the heading part of each table. Can you please translate the content into English and keep the original document format in the tables?"

3) The third step is to carry out MT output assessment both quantitatively and qualitatively using the HOPE metric. For this step, we prepared two sets of triplets (source, reference, MT output) for both ChatGPT and Google MT (Johnson et al., 2017). Afterwards, post-editing was conducted on the reference gold standard in English, resulting in the creation of the **DCC** (Diamond Crisis Corpus) for ITALERT. [2]

3.1 Review of HOPE Metric calculations

HOPE uses error point scores to reflect how much effort is needed to post-edit a machine translation into a correct or gold-standard one. The error points are often annotated segment-by-segment (*e.g.,* a sentence as a segment). In its original formula, HOPE defines each segment as a translation unit (TU), and the Error Point Penalty of that unit (EPPTU) is calculated as the summation of each error type weighted by its severity level:

$$\text{EPPTU} = \sum_{i} Error_i \times Severity(i) \qquad (1)$$

where $Error_i$ is the type of error and $Severity(i)$ is the corresponding point value. Each translation unit is annotated independently from other units.

[2]We offered the reference set because of its availability. The HOPE metric itself does not require the "reference" set because it assumes the annotators understand the source and target language well, so the annotators can use the source to edit the system output, producing a post-edited gold standard.

At the system level, the HOPE metric can be calculated in two different ways. The first way is to sum all penalties for all j translation units, as below:

$$\sum_{TU_j} EPPTU_j = \sum_{i,j} Error_i \times Severity(i) \qquad (2)$$

This metric reflects the system- or document-level effort required to make the translated output correct. The second way is to calculate a similar metric using words as units instead of segments. This approach captures how many (or what percent) of words in the system output fall into the various error categories. In this study we chose the first option, using translation units to measure the system-level scores.

4 Experimental Work

4.1 Experimental Setup

For our experiment, we aimed to test ChatGPT as a representative of Generative AI and compare it with Google Translate as an NMT model, using the ITALERT dataset, as detailed in Table 1.

The evaluation of these two systems consisted of a two-step process. In the first step, we used the HOPE metric to evaluate the outputs. This metric returns both qualitative and quantitative analyses. Error annotation was performed by three professional linguists, native in Italian and proficient in English, all with a Master's Degree in Translation and a Bachelor's Degree in Linguistics. In the second step, post-editing was performed on the reference gold standard to create the DCC corpus. However, this version was not used during the error annotation process, as it will serve as a future resource for further testing and evaluation.

4.2 LLMs vs. NMT: comparative results on the crisis translation task

In this section we present: 1) descriptive statistics on error types and severities; and 2) a qualitative analysis of those errors.

Table 2 compares the performance of each system, listing the number of segments with error scores of 1, 2, 4, and >4, which we call minor, medium, major, and severe.

Error Type	Score	ChatGPT	Google MT
minor	1	104	109
medium	2	73	68
major	4	23	22
severe	>4	30	33
Total		**230**	**232**
Error rate		**0.52**	**0.53**

Table 2: Comparison of segment-level error counts for ChatGPT and Google MT on the crisis translation task. The error rate is computed over 440 total segments.

Examining this, we observe that:

- ChatGPT and Google MT present distinct patterns in segment-level error severity. Notably, ChatGPT produces more medium-severity errors (penalty score = 2) compared to minor ones (score = 1), whereas Google MT shows the opposite trend.
- Both systems exhibit a higher number of severe errors (score > 4) than major errors (score = 4), suggesting that when errors do occur, they often reach high levels of criticality. While the number of major errors is comparable (23 for ChatGPT and 22 for Google MT), Google MT presents slightly more severe errors (33 vs. 30), potentially raising concerns in high-impact applications.
- Overall, the total number of segments with error scores is 230 for ChatGPT and 232 for Google MT, resulting in error ratios of $230/440 = 0.52$ and $232/440 = 0.53$, respectively over 440 test segments. These values indicate that more than half of the evaluated segments contain non-trivial errors, underscoring the need for further system improvements to ensure reliability in sensitive domains such as crisis communication.

Table 3 reports the absolute error counts and percent error per category for ChatGPT and Google MT, based on a total of 611 and 728 annotated errors respectively.

- ChatGPT shows the highest number of errors in STL, TRM, and MIS, with fewer instances in IMP, PRF, UGR, and PRN.
- Google MT also shows the most errors in STL, MIS and IMP, followed by TRM and PRF. No errors were observed in the PRN category.
- The top error categories for both systems are STL, MIS, IMP, and TRM, indicating shared challenges across style, context, and terminology levels.
- The percent errors confirm STL as the most dominant category for both ChatGPT (40%) and Google MT (34%). For ChatGPT, TRM (25%) and PRF (8%) follow, whereas for Google MT, MIS (22%) and IMP (15%) are more prominent.

4.3 Inter-annotator agreement (IAA)

The annotation process was carried out by three professional linguists, native in Italian and proficient in English, all holding a Master's degree in Translation. After drafting a common set of guidelines, the annotators conducted an initial round of annotation to ensure consistency and a shared understanding of the annotation categories. During this phase, borderline cases and ambiguous instances were collected and discussed in a dedicated meeting. As a result of these discussions, the guidelines were refined and updated. In particular, the RAM (Required Adaptation Missing) category was merged into MIS (Mistranslation) to reduce overlap and improve clarity. Furthermore, a decision tree was developed to support the annotators in the classification process and facilitate decision-making for each of the categories.

The final IAA score was computed on a subset corresponding to 10 percent of the entire corpus. This evaluation aimed to assess the reliability and consistency of the annotation process after the consolidation of the guidelines. We measured IAA using well-established reliability measures commonly applied in computational linguistics research (Artstein and Poesio, 2008), drawing on prior work in MT evaluation and annotation reliability (Castilho, 2021). Our metrics include

Model	IMP	TRM	UGR	MIS	STL	PRF	PRN	Total Errors
ChatGPT	72 (11.7)	154 (**25.2**)	11 (1.8)	77 (12.6)	246 (**40.2**)	49 (**8**)	2 (**0.3**)	611
Google MT	111 (**15.2**)	134 (18.4)	19 (**2.6**)	162 (**22.2**)	248 (34)	54 (7.4)	0 (0)	728

Table 3: Absolute error scores for ChatGPT and Google MT in seven categories of errors from the HOPE model. The percent error is shown in parentheses. **Bold** indicates the highest percentage error in each column.

inter-rater agreement (IRR), Cohen's Kappa (Cohen, 1960) for pairwise comparisons, Fleiss' Kappa (Fleiss, 1971) and Krippendorff's Alpha (Krippendorff, 2011) for multi-annotator agreement.

Overall, these metrics confirm a high degree of annotation consistency across systems (Table 4). Both ChatGPT and Google MT reached strong levels of agreement according to multiple IAA metrics. Specifically, Percent Agreement was slightly higher for Google MT (IRR = 92.86) compared to ChatGPT (IRR = 90.48), suggesting that Google MT outputs might have been easier to classify in terms of error presence or absence. Similarly, Fleiss' Kappa and Krippendorff's Alpha confirmed substantial agreement levels for both systems, with Google MT again achieving marginally higher scores ($\kappa = 0.82$, $\alpha = 0.83$) than ChatGPT ($\kappa = 0.78$, $\alpha = 0.79$).

Table 5 presents the pairwise Cohen's Kappa scores for each combination of annotators. The results indicate substantial agreement across all pairs. Interestingly, annotators 1 and 2 showed the most agreement on Google MT ($\kappa = 0.916$), but performed significantly less well on ChatGPT ($\kappa = 0.076$), almost a 20-point difference. This suggests that some annotators encountered significant challenges across different outputs. Overall, IAA was lower on ChatGPT, whereas the higher pairwise agreement on Google MT, combined with its higher IRR scores, may indicate that its outputs were more predictable or less ambiguous in terms of error types *(e.g.* Style, Terminology, Proofreading, etc.), facilitating consistent judgments across annotators.

Model	IRR (%)	Fleiss κ	Kripp. α
ChatGPT	90.48	0.78	0.79
Google MT	92.86	0.82	0.83

Table 4: Inter-annotator coefficient scores for ChatGPT and Google MT: Inter-rater reliability, Fleiss' κ, and Krippendorff's α.

4.4 Evaluations: Levenshtein-perspective

To better understand the differences in string similarity between system outputs and reference translations at both character and word editing levels, we calculate the Levenshtein distance (Levenshtcin, 1966) for both systems against the reference translation on the English side.[3] The Levenshtein Distance (LevDis) measures the similarity or difference between two strings by counting the number of deletions, insertions, and substitutions required to transform one string into the other.

As in Table 6, the overall LevDis from ChatGPT is 11,812 compared to 10,544 from Google MT, across 440 segments. The average LevDis per segment is 29.82 for ChatGPT and 26.62 for Google MT, indicating that ChatGPT outputs are, on average, less similar to the reference strings than Google MT outputs. This is a very interesting outcome with two possible explanations. Assuming our human evaluation is correct and reliable, either LevDis is not a good metric to measure the text similarity, or the reference used has limitations.

Inappropriate metric: It is possible that LevDis is not an adequate or even appropriate metric to measure text similarity, especially semantic-wise, since it only matches the string similarity at the surface level. For instance, phenomena like negation can have a significant impact on language closeness, but are treated with equal weight as any other token in the LevDis calculation. Prior research supports this concern; for example, (Greenhill, 2011) argues that LevDis fails to identify language closeness in the tested data.

Limitations of the reference translation: It is also possible that relying on a single reference translation is limiting, as it may not adequately capture the variability and richness of natural language. Future work should therefore consider multi-reference evaluation settings to better account for this variation and provide a more robust assessment of translation quality.

[3] https://xlinux.nist.gov/dads/HTML/Levenshtein.html

Model	Annot. 1 vs 2	Annot. 2 vs 3	Annot. 1 vs 3
ChatGPT	0.76	0.76	0.84
Google MT	0.92	0.83	0.75

Table 5: Cohen's Kappa scores for ChatGPT and Google MT.

Levenshtein Dis	GPT-ref	Google-ref
Total Distance	11,812	10,544
Avg. dist./seg	29.82	26.62

Table 6: Levenshtein Distance scores comparing System Outputs (ChatGPT and Google MT) against the Original Reference

System	COMET	BLEU
Baseline: GPT	**88.83**	46.29
Google MT	88.73	**50.67***

Table 7: Evaluation results generated with MATEO. * indicates a significant difference with the first row (baseline).

4.5 Automatic Evaluation Metrics

To assess the overall performance of the systems under comparison, we employed two widely-used automatic metrics: BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) using MATEO (Vanroy et al., 2023). Due to missing references, 43 segments were excluded from the evaluation. All reported scores are based on the remaining 397 segments for which gold-standard reference translations were available. This filtering was applied consistently across systems to preserve comparability. All scores were computed with 1,000 bootstrap resampling iterations to estimate 95% confidence intervals, following SacreBLEU-compatible practices.

As shown in Table 7, Google MT significantly outperforms the GPT-based system in terms of BLEU, with a notable difference (50.67 vs. 46.29), indicating a higher degree of n-gram overlap with reference translations. On the other hand, COMET scores, which incorporate both reference and source-side information, are slightly higher for the GPT system (88.8 ± 0.7) compared to Google MT (88.7 ± 0.8). However, the difference is not statistically significant ($p = 0.2258$), indicating that both systems perform comparably on semantic evaluation. These results suggest that while GPT may retain a marginal edge in semantic coherence, Google MT is more aligned with reference outputs at the lexical level.

4.6 Qualitative Assessments of Two Systems

In this section, we categorize the errors made by ChatGPT and Google MT, providing examples based on the HOPE metric analysis presented in the previous section. Next, we summarise key aspects of the crisis domain MT, including the corpus itself and the performance of the translation systems. This is followed by a discussion to provide further insights into the challenges and considerations specific to this domain.

In Table 8, we report examples of stylistic errors found in ChatGPT's translation output. For instance, in the first sentence, ChatGPT translated "seminterrati" using "semi-basements". The term "semi-basements" is confusing and uncommon in English, which may compromise clear communication in a crisis context. In the second example, ChatGPT's use of "it may stop" introduces ambiguity, as "stop" could suggest a pause rather than a malfunction. In contrast, the reference term "get stuck" more effectively conveys the severity of the risk, especially in the context of elevator usage.

Further stylistic issues include the use of "safe spots" instead of "safe places," which feels less consistent with standard emergency communication terminology. Additionally, terms like "prohibit" and "hazards" may lead to confusion, as they are not part of everyday language and may be difficult for the general public to understand. These errors highlight how stylistic inaccuracies can impact the clarity and effectiveness of emergency communication, especially for lay readers.

In Table 9, we provide examples of mistranslations produced by Google MT. One critical error occurs in the translation of the sentence "utilizza un fermo per l'apertura degli sportelli dei mobili," which is rendered as "use a stopper to open the doors of the furniture", failing to convey the intended meaning. Additional errors, such as "closed space" and "order to stay away," are examples of linguistic calques from the Italian language. In

Source	ChatGPT	Google MT	Reference
Non scendere in cantine, seminterrati o garage per mettere al sicuro i beni: rischi la vita.	Do not go into basements, **semi-basements**, or garages to secure belongings: your life is at risk.	Do not go down into cellars, basements or garages to secure your belongings: you risk your life.	Do not enter cellars, basements, or garages to secure possessions; you will risk your life.
Evita l'ascensore: si può bloccare.	Avoid using the elevator: it may **stop**.	Avoid the elevator: it can get stuck.	Avoid the elevator: it may get stuck.
Individua i punti sicuri dell'abitazione dove ripararti in caso di terremoto.	Identify **safe spots** in your home where you can **take cover** during an earthquake.	Identify safe places in your home where you can take shelter in the event of an earthquake.	Identify safe places in your home to take shelter during an earthquake.
L'interdizione della zona rossa potrà durare molto tempo e l'accesso all'area sarà presidiato dalle forze dell'ordine.	Access to the red zone may be **prohibited** for a long time, and the area will be monitored by law enforcement.	The red zone ban may last a long time and access to the area will be supervised by law enforcement.	The red zone exclusion may last for a long time, and access to the area will be controlled by police forces.
Per questo, verifica che la tua casa sia sicura e adotta tutti gli accorgimenti necessari a eliminare le situazioni che possono rappresentare un pericolo.	Therefore, check if your home is safe and take all necessary precautions to eliminate potential **hazards**.	For this reason, make sure your home is safe and take all necessary precautions to eliminate situations that may represent a danger.	Therefore, verify that your home is safe and take all necessary measures to avoid situations that may pose a danger.

Table 8: ChatGPT Style Error Examples

Table 10, we present more error examples from different categories for both systems.

4.7 MT Challenges in Crisis Texts

Within the scope of our investigation, we believe that several challenges related to the crisis corpus and its translation represent valuable topics for discussion.

- *Literal vs context-appropriate translations:* In one example, the source text "Iniziano le operazioni di allontanamento delle persone con particolari necessità di assistenza sociosanitaria" was translated by Google MT as "Operations to **remove people** with particular needs for social and healthcare assistance begin". The phrase "to remove people" introduces a problematic lexical choice. In the context of a natural disaster or emergency, the expected term is "to evacuate," which carries a neutral connotation. By contrast, "to remove people" may imply coercion or force, potentially distorting the communicative intent of the source and undermining trust in the message. This example underscores the importance of context-aware lexical selection in high-stakes scenarios such as crisis communication. Another case is "Non scendere in cantine, seminterrati o garage durante l'alluvione," which was translated as "Do not go down into basements, **semi-basements**, or garages during the flood." The term "semi-basements" is rarely used in English and may confuse readers. These examples demonstrate how overly literal translations can reduce clarity and accessibility, particularly in crisis contexts. Adopting context-sensitive phrasing ensures better clarity and

Source	ChatGPT	Google MT	Reference
Da solo	On your own	**Alone**	On your own
In cucina, utilizza un fermo per l'apertura degli sportelli dei mobili dove sono contenuti piatti e bicchieri, in modo che non si aprano durante la scossa.	In the kitchen, use latches on cabinet doors containing plates and glasses to prevent them from opening during a tremor.	In the kitchen, **use a stopper to open the doors** of the furniture where plates and glasses are stored, so that they do not open during the earthquake.	In the kitchen, secure the cupboard flaps where plates and glasses are stored so they do not open during the earthquake.
Se sei in un luogo chiuso	If you are indoors	If you are in a **closed place**	Indoor
Iniziano le operazioni di allontanamento delle persone con particolari necessità di assistenza sociosanitaria.	The evacuation operations begin for people with specific social and healthcare needs.	Operations **to remove people** with particular needs for social and healthcare assistance begin.	In this phase evacuation operations begin for people with special social and health care needs.
Una volta diramato l'ordine di allontanamento, vai a casa e prepara la valigia.	Once the evacuation order is issued, go home and pack your suitcase.	Once the **order to stay away** has been issued, go home and pack your suitcase.	Once the evacuation order has been issued, go home, and pack your suitcase.

Table 9: Google MT Mistranslation Error Examples

accessibility in critical situations.

- *Impact of passive versus active voice on readers' understanding:* In Table 10, a Google MT Style Error is illustrated where one sentence is translated using the passive voice: "preventive actions that **can be taken**" instead of the more direct and active phrasing: "you can take preventive actions." We believe the choice of voice can significantly influence how recipients perceive and act upon the message. Active voice tends to be more transparent and more engaging, potentially making instructions easier to follow, especially in high-stakes scenarios typical of the crisis domain.

- *Terminology inconsistencies:* Errors in this category are particularly significant in the crisis domain, as they directly affect message accuracy and clarity. Notably, they emerge as the most frequent error type across both systems, as shown in the Table 3. Examples include "hazard" versus "danger," "closed space" versus "indoor," "voids" instead of "sinkholes," and "attention phase" instead of "alert phase". Addressing such discrepancies is essential to ensure precise and actionable communication during crises.

- *Meaning shift:* Some translation outputs from the tested systems result in meaning shifts or changes from the original text, which can have severe implications in the crisis domain. These examples include "false ceiling" *vs*. "suspended ceiling", "removal order" *vs*. "evacuation order".

- *Complex and long sentences:* The source corpus contains numerous long and complex sentences, which can hinder users' ability to process information effectively, leading to reduced actionability. We believe that instructions in crisis communication should be as concise and straightforward as possible for practical use. For instance, the source sentence: "In cucina, utilizza un fermo per

l'apertura degli sportelli dei mobili dove sono contenuti piatti e bicchieri, in modo che non si aprano durante la scossa." (as shown in Table 9), along with its reference, is lengthy and employs complex syntax. Simplifying and splitting such sentences would make them easier to process for both MT systems and end users, enhancing readability and usability in critical scenarios.

- *Register level*: Register plays a crucial role in ensuring that crisis communication is accessible and appropriate for its target users. We recommend using plain or lay language wherever possible to improve accessibility and comprehension. For instance, in the stylistic error highlighted in Table 10, both systems use the phrase "prohibit it," which is more formal, instead of the simpler and more commonly used lay term "ban it." Using plain language ensures that messages are accessible and easy to understand, particularly in high-pressure situations where clear communication is vital.

5 Conclusion and Future Work

In this study, we investigated the performance of LLMs and NMT systems in translating crisis-related texts. The evaluation was conducted using 440 segments from eight subdomains, with data sourced from the national communication campaign website *Io non Rischio*. ChatGPT-4o and Google Translate were selected as representatives of Generative AI and stand-alone NMT systems respectively, and were evaluated using a human-centric evaluation framework.

Errors from each system were categorised using the default 7 error types (merged from 8) from the HOPE metric, with a revised severity mapping, adjusted to account for the sensitivity of the crisis domain. The findings reveal that both systems share common error types but differ in their rankings. ChatGPT showed a high incidence of Style and Terminology errors, while Google MT was characterised by a greater presence of Mistranslation, Impact, Terminology, and Style issues. Importantly, both systems produced a non-negligible amount of severe and major errors, despite the predominance of minor and medium-level issues. The number of segments with severity ratings above 4 was slightly higher in Google MT outputs than in those of ChatGPT, indicating a greater incidence of critical errors. As seen in the qualitative analysis, several of these high-severity errors in Google MT translations had the potential to significantly distort the intended meaning and undermine the actionability of the messages. Interestingly, automatic evaluation metrics appear to diverge from human error analysis findings. While BLEU scores show a clear advantage for Google MT, indicating stronger surface-level fidelity, COMET scores are only marginally higher for the GPT system, suggesting comparable semantic adequacy. This trend aligns more closely with human judgments: HOPE-based error annotation reveals that Google MT's surface-level advantage does not correspond to improved quality in critical cases, as human annotators identified 728 errors in Google MT outputs, compared to 611 in those produced by ChatGPT. This discrepancy reflects limitations of using merely quantitative metrics in capturing context-sensitive, high-impact errors, and highlights the importance of complementary human-centric evaluations, especially in high-stakes scenarios (Hajek et al., 2024), as well as the need of developing domain specific automatic metrics in the future, e.g. for crisis translation.

In addition to the error analysis of current LLMs and NMT systems, using the HOPE framework, we also produced the Diamond Crisis Corpus (**DCC**), a new post-edited reference set derived from the ITALERT dataset.

Since ITALERT represents the first MT corpus on Italian crisis translation, we plan to add new subdomains, such as *public health emergencies* from the healthcare domain (Han et al., 2024). Other authoritative sources from which to extract the texts include Médecins Sans Frontières[4], The International Red Cross[5], and The United Nations Office for Disaster Risk Reduction[6]. We also plan to calculate the Levenshtein distance between the systems' outputs and the DCC corpus, to investigate differences in findings compared to those obtained using the original "gold" corpus. This analysis will help determine whether the post-edited corpus enhances the evaluation of translation performance.

Future work will explore in-domain training to address the challenges of context disambiguation and terminology management (Kirchhoff et al., 2011) and raise awareness of a responsible use of translation technology in high-stakes settings.

Finally, while we acknowledge that resources

[4] https://www.msf.org
[5] https://www.icrc.org/en
[6] https://www.undrr.org

comparable to ITALERT are not yet widely available, this work represents an initial step toward addressing the underexplored area of multilingual crisis communication (Cadwell et al., 2024), hoping it will serve as a foundation for future research and resource development in this emerging field.

Acknowledgments

This work was partially supported by the PhD programme in Humanities and Technologies funded by the University of Macerata under D.R. No 253/2023. We thank Professor Federico Federici from UCL for his valuable suggestions on the selection of texts for our corpus. We thank Kung Yin Hong (Kenrick) for helping with the HOPE metric and Levenshtein scores in the earlier stages of this work. We thank Willemijn Klein Swormink for the valuable discussion on a more interactive, visualised, and insightful database to build for the ITALERT project. We thank Serge Gladkoff, the CEO of Logrus Global LLC, for valuable advice on the manuscript. We thank Argentina Anna Rescigno and Antonio Castaldo for their contributions to the annotation process and for the brainstorming sessions, which helped us refine the annotation guidelines for the HOPE metric.

Sustainability statement

In this study, we conducted an evaluation of current LLMs and NMT systems for translation quality assessment within the crisis domain, without performing any training or fine-tuning. The computational requirements were minimal, as the evaluation involved translating only 440 segments.

References

Amalia Amato and Letizia Cirillo. 2024. Mediating english as a lingua franca for minority and vulnerable groups-introduction. *MEDIAZIONI*, (41):1–7.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Patrick Cadwell, Sharon O'Brien, and Ezequiel DeLuca. 2019. More than tweets: A critical reflection on developing and testing crisis machine translation technology. *Translation Spaces*, 8(2):300–333.

Patrick Cadwell, Sharon O'Brien, Aline Larroyed, and Federico M Federici. 2024. A crisis translation maturity model for better multilingual crisis communication. *INContext: Studies in Translation and Interculturalism*, 4(2):136–165.

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.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Yang Cui, Lifeng Han, and Goran Nenadic. 2023. MedTem2.0: Prompt-based temporal classification of treatment events from discharge summaries. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 160–183, Toronto, Canada. Association for Computational Linguistics.

Federico M. Federici. 2016. *Mediating Emergencies and Conflicts*. Palgrave Macmillan, Houndmills.

Federico M. Federici and Patrick Cadwell. 2018. Training citizen translators: Design and delivery of bespoke training on the fundamentals of translation for new zealand red cross. *Translation Spaces*, 7(1):23–43.

Selim Fekih, Benjamin Minixhofer, Ranjan Shrestha, Ximena Contla, Ewan Oglethorpe, Navid Rekabsaz, et al. 2022. Humset: Dataset of multilingual information extraction and classification for humanitarian crises response. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4379–4389.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Serge Gladkoff and Lifeng Han. 2022. HOPE: A task-oriented and human-centric evaluation framework using professional post-editing towards more effective MT evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 13–21, Marseille, France. European Language Resources Association.

Simon J Greenhill. 2011. Levenshtein distances fail to identify language relationships accurately. *Computational Linguistics*, 37(4):689–698.

John Hajek, Yu Hao, Ambrin Hasnain, Anila Hasnain, Ke Hu, Maria Karidakis, Rachel Macreadie, Anthony Pym, and Juerong Qiu. 2024. Understanding and improving machine translations for emergency communications.

Lifeng Han, Serge Gladkoff, Gleb Erofeev, Irina Sorokina, Betty Galiano, and Goran Nenadic. 2024. Neural machine translation of clinical text: an empirical investigation into multilingual pre-trained language models and transfer-learning. *Frontiers in Digital Health*, 6:1211564.

Lifeng Han, Alan Smeaton, and Gareth Jones. 2021. Translation quality assessment: A brief survey on manual and automatic methods. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 15–33, online. Association for Computational Linguistics.

ISTAT. 2023. Nuovo picco delle presenze turistiche. Accessed: 2024-11-27.

ISTAT. 2024. Intensi flussi di immigrazione straniera, in lieve ripresa mobilità interna ed espatri. Accessed: 2024-05-28.

Wenxiang Jiao, Wenhui Wang, Jitong Huang, Xu Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint*, arXiv:2301.08745(1):1–10.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Katrin Kirchhoff, Anne M Turner, Amittai Axelrod, and Francisco Saavedra. 2011. Application of statistical machine translation to public health information: a feasibility study. *Journal of the American Medical Informatics Association*, 18(4):473–478.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

VI Lcvenshtcin. 1966. Binary coors capable or 'correcting deletions, insertions, and reversals. In *Soviet Physics-Doklady*, volume 10.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2024. Towards explainable evaluation metrics for machine translation. *Journal of Machine Learning Research*, 25(75):1–49.

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*.

Michael K. Lindell, Ronald W. Perry, Carla Prater, and William C. Nicholson. 2006. Fundamentals of emergency management. page 485. FEMA, Washington, DC, USA.

Arle Lommel, Serge Gladkoff, Alan K Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, et al. 2024. 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: User Track)*, pages 75–94.

Sharon O'Brien and Patrick Cadwell. 2017. Translation facilitates comprehension of health-related crisis information: Kenya as an example. *JoSTrans: The Journal of Specialised Translation*, 28:23–51.

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.

Stefano Perrella, Lorenzo Proietti, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. Beyond correlation: Interpretable evaluation of machine translation metrics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20689–20714, 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.

Bram Vanroy, Arda Tezcan, and Lieve Macken. 2023. MATEO: MAchine Translation Evaluation Online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 499–500, Tampere, Finland. European Association for Machine Translation.

A Error Examples

We list in Table 10 additional error examples from Google MT and ChatGPT.

Source	ChatGPT	Google MT	Reference
Valuta bene se mettere al sicuro l'automobile o altri beni: può essere pericoloso.	Carefully consider if securing your car or other belongings is necessary, as it can be dangerous.	Consider carefully whether to secure your car or other **property**: it can be dangerous.	Carefully evaluate whether to secure your car or other possessions; it can be dangerous.
In questa fase, quando viene diramata un'allerta, è possibile compiere alcune azioni preventive per ridurre il rischio.	During this phase, when an alert is issued, you can take some preventive actions to reduce risk.	At this stage, when an alert is issued, there are some preventive actions that **can be taken** to reduce the risk.	You can take preventive actions to reduce the risk when an alert is issued.
Fai un elenco delle cose essenziali da portare con te (documenti, medicine, occhiali da vista, ecc.).	Make a list of essential items to bring with you (documents, medications, **glasses**, etc.).	Make a list of essential things to bring with you (documents, medicines, **glasses**, etc.).	Make a list of essentials to bring (documents, medicines, eyeglasses, etc.).
Prima di bere l'acqua dal rubinetto assicurati che ordinanze o avvisi comunali non lo vietino; lava bene i cibi che sono venuti a contatto con le ceneri vulcaniche.	Before drinking tap water, make sure municipal ordinances or advisories do not **prohibit** it; thoroughly wash food that has come into contact with volcanic ash.	Before drinking tap water, make sure there are no ordinances or notices **prohibiting** it; wash any food that has come into contact with volcanic ash thoroughly.	Before drinking water from the tap, make sure that municipal ordinances or notices do not **ban** it; wash well any food that has come into contact with volcanic ashes.

Table 10: Additional Error Examples from ChatGPT and Google MT

Optimising ChatGPT for creativity in literary translation: A case study from English into Dutch, Chinese, Catalan and Spanish

Shuxiang Du[1†], **Ana Guerberof Arenas**[1†], **Antonio Toral**[2†]
Kyo Gerrits[1], **Josep Marco Borillo**[3]
[1]Centre for Language and Cognition, University of Groningen
[2]Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant
[3]Departament de Traducció i Comunicació, Universitat Jaume I
sophie321du@gmail.com

Abstract

This study examines the variability of ChatGPT's machine translation (MT) outputs across six different configurations in four languages, with a focus on creativity in a literary text. We evaluate GPT translations in different text granularity levels, temperature settings and prompting strategies with a Creativity Score formula. We found that prompting ChatGPT with a minimal instruction yields the best creative translations, with "Translate the following text into [TG] creatively" at the temperature of 1.0 outperforming other configurations and DeepL in Spanish, Dutch, and Chinese. Nonetheless, ChatGPT consistently underperforms compared to human translation (HT). All the code and data are available at https://github.com/INCREC/Optimising.

1 Introduction

The intersection of artificial intelligence (AI) and creativity in the domain of translation presents a fascinating and challenging field for research. Even if the development of machine translation (MT) technologies, especially through the advent of Large Language Models (LLMs) like ChatGPT, has reshaped the landscape of the language industries, there remains a notable gap in the creative capacities of MT outputs in comparison to that of professionals (Karpinska and Iyyer, 2023). This type of translation, often applied to literary texts, requires not just the accurate conveyance of meaning but also the preservation of style, tone, and creative nuances inherent in the source text to create an effect on the reader that is not purely information driven. Since new models offer a dialogic capacity, we explore in this paper the best set of variables to generate the most creative translations using ChatGPT.

We investigate various configurations of ChatGPT, including different text granularities, temperature settings, and prompting strategies, alongside a comparison with translations by neural machine translation (NMT) systems, and human translations as references. The primary aim is to evaluate how these configurations impact the creativity and quality of the translations produced, using both a manual creativity scoring system and automatic evaluation metrics. The experiment involves translating a short science fiction story by Kurt Vonnegut, "2BR02B", from EN (English) to ZH (Chinese), NL (Dutch), CA (Catalan), and ES (Spanish). The translations are manually annotated to assess creative shifts (CSs) and errors, providing a detailed analysis of how ChatGPT in different configurations handles the nuanced demands of literary translation. Thus, central to this research are the questions:

RQ1: What is the variability in MT outputs from ChatGPT under different settings?

RQ2: What is the optimal prompting setting for the most creative MT output using ChatGPT?

2 Related Work

LLMs like ChatGPT have demonstrated promising performances in natural language processing tasks (Kalyan, 2023). LLMs such as IOL-Research, Unbabel Tower 70B, and Claude-3.5-Sonnet are the top performing MT systems submitted to the last edition of WMT's general translation task (Kocmi et al., 2024). ChatGPT for MT has demonstrated promising applications to help users translate specific contents or entire documents, especially between high-resourced languages (Jiao et al., 2023, Hendy et al., 2023). However, whether it outperforms NMT systems or commercial MT systems is still under debate (Kalyan, 2023).

The research community has investigated the effectiveness of ChatGPT for MT in different aspects.

[†]Equal contribution

Proceedings of Machine Translation Summit XX Volume 1, pages 578–591, June 23-27, 2025

Gao et al. (2024) focused on developing advanced prompting strategies by including additional information like task, domain, and syntactic information like PoS (parts of speech) tags. The researchers tested the language pairs English ↔ Spanish, English ↔ French, and Spanish ↔ French in the domains of news, e-commerce, social, and conversational in a sentence level. They concluded that including appropriate information about the input text in the prompt, such as specifying translation task or context domain, can improve the performance of ChatGPT. ChatGPT has a higher BLEU (Papineni et al., 2002) score in four out of the six language pairs when compared to Google Translate (GT) and DeepL Translate (DeepL) with their proposed advanced prompting strategies.

In terms of text granularity levels, Wang et al. (2023) examined the performances of ChatGPT for document-level translation, covering three language pairs (Chinese ⇒ English, English ⇒ German, and English ⇒ Russian) in seven domains (news, social, fiction, Q&A from an online forum, TED, Europarl, and subtitle). The researchers reported that ChatGPT does well when the sentences in the document are combined and given at once to the model. With this prompting strategy, it exhibited better performances than commercial MT systems according to human evaluation and also outperformed most document-level NMT methods in terms of d-BLEU scores.

Temperature is a hyperparameter in LLMs that regulates the randomness in text generation by adjusting the probability distribution of potential next words (Peeperkorn et al., 2024). Decoding with higher temperatures displays greater linguistic variety, while low values tend to generate grammatically correct and more deterministic text (Ippolito et al., 2019). Peng et al. (2023) explored the impact of temperature, task, and domain information on the translation performance of ChatGPT. In the translation of English, Chinese, German, and Romanian of biomedicine, news, and e-commerce texts, the study showed that ChatGPT performance degraded with an increase in temperature in terms of both BLEU and COMET (Rei et al., 2020) scores, and hence it was recommended to use a lower temperature (recommended is 0 for their test set). Additionally, including task and domain information in the prompt enhanced the translation performance of ChatGPT consistently for both high- and low-resource languages in their research.

As MT technology advances, there is growing interest in exploring how well these systems can handle the complexities of literary translation. With respect to LLMs, Karpinska and Iyyer (2023) evaluated the performance of ChatGPT in translating literary paragraphs across 18 linguistically diverse language pairs. The authors experimented with three different prompting strategies, namely translating sentence by sentence in isolation, translating sentence by sentence in the presence of the rest of the paragraph, and translating the entire paragraph at once. According to human evaluation, when translating entire paragraphs, ChatGPT produced translations of significantly higher quality compared to other strategies and commercial systems. However, critical errors such as content omissions still occur. The findings suggest that while ChatGPT can leverage larger context units like paragraphs to enhance translation quality, this is yet not sufficient on their own for high-stakes applications like literary translation where nuanced understanding and stylistic consistency are crucial.

The challenges of literary MT lie not only in the performance of the systems but also in the evaluation of the results. Fonteyne et al. (2020) provided an in-depth evaluation of the quality of a novel translated by NMT from English ⇒ Dutch. Unlike traditional sentence-level evaluations, this study emphasized the importance of document-level analysis to better assess the coherence and cohesion of translated texts, which are crucial in literary translations. It utilized an adapted version of the SCATE error taxonomy (Tezcan et al., 2017), which considers errors at both the sentence and document levels. Again, the findings suggested that while NMT can produce a substantial portion of error-free translations, significant errors remain, particularly with complex elements like style and coherence that are vital to literary texts. Therefore, it is important to consider metrics other than error annotation when evaluating literary texts.

In the studies of ChatGPT for translation, most evaluations focus on automatic metrics like COMET and BLEU, while the specific aspect of creativity has hardly been touched upon. This could be because creativity in translation is hard to measure. Bayer-Hohenwarter (2009) proposed a framework for assessing translational creativity based on the concepts of novelty, acceptability, flexibility, and fluency. Novelty in translation is characterized by three main aspects: exceptional performance that significantly surpasses routine translation activities, uniqueness or rarity within a specific cor-

pus of translations, and non-obligatory translational shifts that indicate a high level of translator engagement and creativity. Acceptability is defined as "skopos adequacy" (Bayer-Hohenwarter, 2009, 2). This emphasizes that a creative translation must not only be innovative but also appropriate and useful within the context for which it is intended. These novelty and acceptability aspects of the framework are largely adopted in this research.

Bayer-Hohenwarter (2011) defines creative shifts as transformative operations in translation that deviate from the direct replication of the source text. These shifts are categorized into three types: abstraction, where translators generalize specific details from the source; modification, which involves alterations to better suit the cultural or contextual needs of the target text; and concretization, where translators add specific details not explicitly mentioned in the source text. Bayer-Hohenwarter proposed a systematic methodology to measure creativity in translation by identifying and analyzing these creative shifts. She defined specific "units of analysis" within the texts, identifying both "creativity units" (requiring high problem-solving capacity) and "routine units" (relatively straightforward translation tasks). The results were quantified by calculating the proportion of creative shifts versus literal reproductions. The study also examined the relationship between the frequency of creative shifts and the overall quality (acceptability) of the translations. There was a general trend suggesting that translators who produced more creative shifts also produced higher-quality translations. However, this was not a strict correlation, as some creative shifts led to errors, particularly among less experienced translators.

Guerberof-Arenas and Toral (2020, 2022) created a formula (see section 3) to quantify creativity in translations, offering a measurable way to assess and compare the creative output of different translation modalities, including MT. Their study involved the translation of literary texts from English to Catalan and Dutch. The texts were translated by professionals, post-edited by professionals, and machine translated. By applying a creativity score to the translations, they found MT outputs to be less creative than professional translations and that they limited the translator's creativity in post-editing. The quantification framework they established is used in this research.

In their Master thesis Du (2024) use this creativity index in an evaluation of ChatGPT translations of a literary text in the English ⇒ Chinese translation direction. They investigated different set-ups of ChatGPT including levels of text granularities, different temperatures, prompting strategies, and few-shot prompting. The findings indicated that the quality and creativity of ChatGPT translations vary across these configurations. The best setting in their study was a document-level translation with a temperature of 1.0 and a direct prompt to be more creative. In this paper, we replicate the experiment with more languages and a more in-depth analysis.

3 Methodology

In this section, we explain the source text (ST) used, how the target texts (TTs) were generated in the different phases of our experimentation, as well as the data annotation and analysis process.

3.1 Source Text

The study utilized a curated dataset comprising different translations of a short science fiction story by Kurt Vonnegut: *2BR02B*[1] (Vonnegut, 1999). The story is a short science fiction piece set in a future society where aging has been cured and the population is strictly controlled to remain at forty million. Individuals must volunteer for death to allow new births. It revolves around a family about to give birth to three kids and therefore in need of three volunteers to die.

This story was selected for three reasons: A) we have an existing corpus of annotations on the units of creative potential in the story (Guerberof-Arenas and Toral, 2022), B) to our knowledge it has not been translated into the target languages to date[2] and therefore we assume it has not been used in the training data of ChatGPT, and C) it requires a high level of translation creativity.

The story was processed in Python to be broken into separate paragraphs. The text overall contains 123 paragraphs, 234 segments and 2548 words. There are 185 units of creative potential (UCP) in total, annotated by two experienced translators and researchers in the previous study (Guerberof-Arenas and Toral, 2022). These are units in the ST that are expected to require translators to use problem-solving skills, as opposed to those that are regarded as routine units with little creative potential (Bayer-Hohenwarter, 2011).

[1] https://www.gutenberg.org/ebooks/21279

[2] Not found in the Unesco Translationum database https://www.unesco.org/xtrans/bsform.aspx nor on National Library of China https://www.nlc.cn/web/index.shtml

3.2 Target Text

For the target text (TT), we used the model gpt-4o-2024-08-06 with the ChatGPT API[3] to translate the text into ZH, NL, ES and CA. This version is chosen for three reasons: A) it was the latest stable model of ChatGPT when we started our experimentation, thus representing state-of-the-art performance, B) according to OpenAI[4] , this version performs better on text in non-English languages, C) in terms of data training, the cost of this version is relatively lower and the speed is faster when compared to ChatGPT-4.

Due to limited capacity and the exploratory nature of this experiment, we decided to annotate a subset of the text. We selected a series of UCPs that were previously singled out by two annotators in the ST to ensure a better representation of the creative potential of this text. In the end, 54 UCPs, present in 48 separate sentences with a total of 602 words were selected for the annotation task in the TT. To prepare the sentence-aligned files, we manually post-processed the text by extracting the 48 sentences in each translation.

Each translation of the sentences in the TT was manually annotated for a detailed comparative analysis of creativity across different translations. The annotators were four of the researchers that are experienced translators or have a language related Master degree in the selected language combinations: there was therefore one annotator per language combination.

As the baseline of the study, DeepL has been chosen to compare with ChatGPT. The reason is that in the preliminary experiment (Du, 2024) it offered a more pleasant-to-read translation than other NMT systems like Google Translate. Since DeepL is not available for CA, we used two popular NMT systems for this target language: Softcatalà's *Traductor*[5] and Google Translate.[6]

3.3 Data Collection

In this experiment, we try a range of text granularities, temperature settings, and zero-shot prompting strategies based on Du (2024) master project to generate translations with ChatGPT. The experiments and annotations were conducted between October 2024 and January 2025. Figure 1 shows an overview of the NL and ZH workflow process as an example. The workflow in each phase slightly differs for the other two languages (CA and ES).

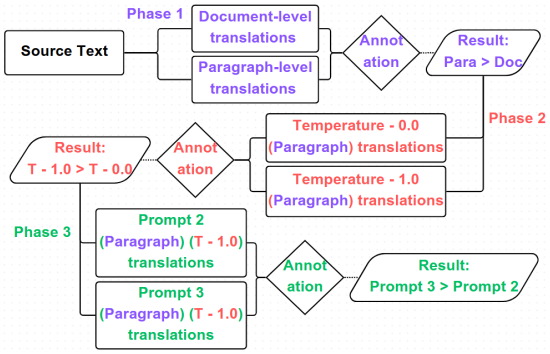


Figure 1: Workflow for ZH and NL

3.3.1 Phase 1. Text Granularity

The variable in the first phase is text granularity. We translated the text at both paragraph level (setting 1a) and document level (1b). At the paragraph level, we entered the same prompt for each paragraph in the story, with each request done separately to avoid context interference. At the document level, we entered the same prompt followed by the entire story in one request. It is worth noting that 14 of the 48 sentences involved in the evaluation process were single-sentence paragraphs.

The prompts used in phase 1 are:

Prompt 1 (1a): "Translate into [TG]: [Input]"

Prompt 1 (1b): "Translate the following text into [TG]: [Input]"

On the one hand, having the whole document available offers more context, which should be useful for its translation. On the other, recent research has shown that translation performance decreases with the length of the input text (Peng et al., 2024).

After the evaluation of the creativity score for these two outputs (see Section 4.1), we proceeded with the better granularity method for each target language to further experiment with different temperatures and prompts. Namely, in CA and ES, the document-level translations were assessed to be better, thus the following experiment in CA and ES were conducted at document level (1b), while the opposite was the case for ZH and NL (1a).

3.3.2 Phase 2. Temperature

The temperatures selected are 0.0 (2a) and 1.0 (2b). In the API setup, ChatGPT's temperature can range from 0.0 to 2.0. The default temperature setting of

[3]Version 1.54.4, i.e. the latest when we started out experiments.

[4]https://openai.com/index/hello-gpt-4o/

[5]https://www.softcatala.org/traductor/

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

ChatGPT is officially stated to be set to 1.0[7]. However, according to our experience the default value seems to be 0.0. Namely, we used the automatic evaluation metric chrF (Popović, 2015) to examine how similar the translations produced using different temperature values are to the translations produced in phase 1, which used the default setting for temperature. For NL, for example, the chrF result decreased as the temperature went up: 88 (temperature=0.0), 86 (0.5), 83 (1.0), 82 (1.1) and 81 (1.2). This means that ChatGPT is not deterministic, even with temperature 0.[8] We then chose the value 0 in phase 2 to see the effect of non-determinism.

On the other hand, the higher the temperature is, the more creative the text is expected to be. However, the highest value we chose was not the maximum offered by the API (2.0) but 1.0. This is because we noticed that at higher temperature values, there are more instances of "word vomit" in the output which makes the text incoherent and impossible to read.[9] Therefore, at temperature 1.0 the system is most likely to generate more creative content while not suffering from word vomit.

The same prompt as in phase 1 was used and we proceeded with the best temperature setting after evaluation (see Section 4.2). For ES, NL, and ZH, we proceeded with temperature 1.0 (2b), while for CA we proceeded with 0.0 (2a).

3.3.3 Phase 3. Prompting Strategies

The zero-shot prompting strategies we designed included prompting with the specific domain information, i.e. author and genre (3a), and prompting with direct instructions to generate creative outputs (3b). The final prompts are as follows:

- **Prompt 2** (more info about genre and author, 3a): Translate the following text into [TG] taking into consideration that this is (from) a science fiction story by Kurt Vonnegut: [input]

- **Prompt 3** (request of creativity, 3b): Translate the following text into [TG] creatively: [input]

[7] https://platform.openai.com/docs/api-reference/chat/create

[8] If it was deterministic, then chrF's score when comparing phase 1's translation with phase 2's translation with temperature=0 would have been 100,

[9] We tried three values of temperature higher than 1.0 (1.1, 1.2 and 1.5) and noticed severe issues with values 1.1 (CA), 1.2 (ES), 1.5 (NL and ZH). We speculate that the reason why ChatGPT has issues in CA and ES at a lower temperature values than NL and ZH is because for the former the document is translated at once.

3.4 Data Annotation

Following data collection, each sentence of each translation was manually annotated in terms of acceptability and novelty, as discussed in section 2. The annotators were blind to the specific setting they were evaluating.

Acceptability was measured according to the number and severity of errors in the TTs based on the harmonized DQF-MQM Framework (Lommel et al., 2014). The severity of each error was marked as: Neutral (0 points for repeated errors or preferences), Minor (1 point), Major (5 points), and Critical (15 points). Minor refers to errors that do not lead to loss of meaning and do not confuse or mislead the reader but are noticeable, hence they decrease stylistic quality, fluency or clarity, or make the content less appealing. Major refers to errors that may confuse or mislead the reader or hinder the understanding of the text due to significant change in meaning or because errors appear in a visible or important part of the content. Critical refers to errors that may misrepresent or damage the reputation of the author or publishing house, causes the text to stop working as a literary artefact and affect the communicative flow, or if the language is perceived as offensive (when unintended), but also, if the text departs from the source text in such a significant way that has a large impact on the understanding of the entire story.

For example: the title of the story, "2BR02B", is a play on words on the famous quote *To be or not to be* in Shakespeare's *Hamlet*. If left in English in ES and CA, the understanding of the entire story is compromised, and it is therefore considered a Critical error. If the word *business* is translated literally in a context where it does not refer to a commercial activity but to a person's concern, then this can be considered a Major error because the entire text is understood albeit with certain difficulties. Finally, a spelling mistake would be considered Minor.

For novelty, the translation solutions to the UCPs selected were annotated in the TTs. All translations that deviate from the ST that are neither the exact reproduction of the ST nor an omission nor an error count as CS and are classified in the following manner: Abstraction refers to instances when translators use more vague, general or abstract solutions. Concretization refers to instances when the TT evokes a more explicit, more detailed, and more precise idea or image. Modification refers to instances when translators use a different solution in

the TT (e.g. express a different metaphor without the image becoming more abstract or concrete).

For example: if the title of the story, *2BR02B*, is translated into CA as *C-O-N-O-C*, a play on words that evocates *Ser o no ser*, the standard Catalan phrase, this would be considered a Modification and classified as CSM. While in NL, the title *2BR02B* is left as is, as the standard phrase in Dutch remains *To be or not to be*. This is then considered a Reproduction, and classified as such. As this exemplifies, Reproductions are not errors by default, although some UCPs that are not translated might be considered as containing an error.

In the process of annotation, the translation of the 54 UCPs was assessed. For each translated UCP, the annotator decided if the resulting TT was a CS, an omission (O), a reproduction (R), or if it was impossible to classify (E). The CSs were further classified into abstraction (CSA), concretization (CSC), and modification (CSM). Each of the 48 sentences were annotated for errors according to the severity criteria described. The total number of CSs and Error points was used for the creativity index, introduced next.

3.5 Data Evaluation

Acceptability and novelty are combined into a single score using the creativity index (CI) formula:

$$\text{CI} = \left(\frac{\#CSs}{\#UCPs} - \frac{\#\text{error points}}{\#\text{words in ST}} \right) \times 100$$

The index considers both novelty (CSs) and appropriateness or acceptability (errors), enabling a quantifiable comparison between different translation modalities (Guerberof-Arenas and Toral, 2020, 2022).

Apart from the creativity index, we used a number of automatic evaluation metrics (AEMs): BLEU, chrF, TER (Snover et al., 2006), COMET and COMET-Kiwi (Rei et al., 2022). The first three are string-based[10] while the last two are based on multilingual language models.[11] Another distinction is that the first four evaluate a translation with respect to a reference translation (see Section 3.6),[12] while the last one does so with respect to the source text. Since we do not have a reference translation for ZH, only COMET-Kiwi was used.

[10]We compute them with sacrebleu 2.5.1

[11]We used models wmt22-comet-da and wmt22-cometkiwi-da, respectively.

[12]COMET takes into account also the ST.

3.6 Human Reference

Since this experiment utilizes a dataset from the Guerberof-Arenas and Toral (2022) project, we had access to translations created by professionals in EN⇒CA, EN⇒NL and EN⇒S,[13] but unfortunately not for EN⇒ZH. Table 1 shows the scores for the selected UCPs for these languages.

	# CSs	# Errors	Error Points	CI
ENCA	21	2	2	40
ENES	22	6	6	40
ENNL	29	17	25	50

Table 1: Creativity Index in Human Reference

The results for ENCA and ENES were annotated by a professional literary translator, while the ENNL was annotated by a different one for that language pair, and this could account for the differences in judgement, although, of course, this could also mean that there are differences in the quality provided by the translators. One aspect to note here is that while annotating the UCPs, the reviewers also remarked that the entire segments contained other CSs. For example, in ES and CA, the translators changed the name of the characters to be able to create meaningful play on words that were present in the ST.

4 Results

The following subsections contain the results obtained in each of the phases explained in the methodology. Detailed annotations per language are provided in Appendix A.

4.1 Phase 1. Text Granularity

Table 2 shows the results for phase 1, ChatGPT outputs at paragraph (1a) and document level (1b) were compared.

In this instance, the best solutions for ES and CA are at document level, as we would expect since the context of the sentences is considered. However, for NL and ZH the best performance is at paragraph level, mainly due to error points.

In the case of NL, the version at document level included more grammatical errors (such as missing articles *van drieling* ("of triplet"), incorrect subject-verb agreement, e.g. *wat je zaken was* ("What your

[13]The Spanish translation was not analyzed in the previous project, but was translated by the researcher to be used as a reference. This version was then annotated for errors by a professional literary translator.

	Paragraph (1a)				Document (1b)			
	# CSs	# Errors	Error Points	Score	# CSs	# Errors	Error Points	Score
ENCA	**5**	**51**	199	-23.80	**5**	**51**	**158**	**-16.99**
ENES	6	59	253	-31.00	**10**	**52**	**219**	**-18.00**
ENNL	**15**	**61**	**108**	**9.84**	12	68	174	-6.68
ENZH	**11**	**51**	**187**	**-10.69**	10	56	298	-30.98

Table 2: Creativity score for two different text granularities: paragraph and document. The best score per language and criterion is shown in bold.

business is", where *zaken* is plural but *was* singular), hallucinations ("formidable" became *ontslagbaar*, a non-existent word, meaning something like *unfireable*), and typos than the version at paragraph level. The paragraph level version does lack consistency at times ("orderly" is translated differently three times), but still has fewer errors.

For ZH, the document-level translations tend to make more grammatical and factual errors, too, especially towards the end of the document. For example, "sheave-carrier" is translated to " 运载屁股的人 (ass-carrier)", which is a critical error that disrupts the narrative significantly. "He was seven feet tall" is translated to " 两个高大的人 (two tall men)", perhaps in an attempt to convert seven feet to two meters. "He said to her as she fell" is translated to " 他对她说，落 (he said to her, falls)" and is not coherent in the target language. Such examples suggest that ChatGPT tends to perform progressively worse for ZH as it processes the whole document. This is in line with previous findings by Wang et al. (2024) that LLMs demonstrate short-comings in long-text translations, and their performance diminishes as document size increases.

For CA, the difference (in quantitative terms) between the paragraph and the document levels is largely accounted for by the fact that, in the latter, all the fanciful sobriquets[14] for an institution (the Federal Bureau of Termination) are translated, whereas at paragraph level only 6 (out of 14) are. In other respects, differences between the two CA versions are not that pronounced.

For ES, the paragraph and document level translations are not that dissimilar quantitatively. However, the document level resolves certain translations problems better. For example, the expression "seven feet tall" is converted at document-level into meters while it remains in feet at sentence level, and "trick telephone number" is translated as *número de teléfono trampificado* which does not exist as a term, while the document-level uses *número de teléfono con truco* that is correct in Spanish.

4.2 Phase 2. Temperature

Table 3 shows the results of ChatGPT outputs when the temperature was set at 0.0 (2a) and at 1.0 (2b).

The best performance for ES, NL and ZH are at a temperature of 1.0, but for CA the best output is at temperature 0.0. For most languages, a temperature value of 1.0 outputs more CSs but also more errors–only in ES does a temperature of 0.0 have more errors–as was expected.

In NL, for instance, the output at temperature 1.0 translates "triplets" as *drieën* (*threes*)–this is more creative and it could work in some contexts, but not when talking about three babies born at the same time. Still, weighing the CSs against the errors in the creativity index reveals that a temperature of 1.0 has a better output for ES, NL and ZH, despite the errors in the last two.

The general trend is observable for CA too–a higher temperature yields both more CSs and more errors. What sets CA apart is that the higher number of CSs does not compensate for the number of errors because of their severity. At temperature 0.0, for example, "Chicago Lying-in Hospital" is adequately translated, whereas at temperature 1.0 the "Lying-in" segment is left untranslated. Other segments are translated in both settings, but the rendering provided at temperature 1.0 is not acceptable. For example, "Kiss this sad world toodle-oo" is translated as *donaré adéu* ('I will give goodbye'), a collocation that does not exist in CA. Also, "Good gravy", used as an interjection, is adequately translated at temperature 0.0 and wrongly rendered as *Bona sort* ("Good luck") at 1.0.

4.3 Phase 3. Prompting Strategies

Table 4 shows the results for ChatGPT outputs when prompting with more information about

[14]These are nicknames given to the gas chambers in this dystopian world, e.g. Weep-no-more, Good-by, Mother or Easy-go

	T-0.0 (2a)				T-1.0 (2b)			
	# CSs	# Errors	Error Points	Score	# CSs	# Errors	Error Points	Score
ENCA	4	**57**	**200**	**-25.82**	**6**	69	248	-30.08
ENES	8	55	216	-21.00	**9**	**50**	**164**	**-11.00**
ENNL	11	**49**	**99**	3.93	**12**	52	108	**6.13**
ENZH	11	**45**	**155**	-5.38	**14**	51	165	**-1.48**

Table 3: Creativity score for two different temperature values: 0.0 and 1.0.
The best score per language and criterion is shown in bold.

	Prompt 2 (3a)				Prompt 3 (3b)			
	# CSs	# Errors	Error Points	Score	# CSs	# Errors	Error Points	Score
ENCA	**4**	57	204	-26.48	**4**	**55**	**202**	**-26.15**
ENES	12	57	244	-18.31	**13**	**43**	**166**	**-3.50**
ENNL	10	43	89	3.73	**18**	**30**	**76**	**20.71**
ENZH	14	39	171	-2.48	**15**	**37**	**161**	**1.03**

Table 4: Creativity score for two different prompting strategies.
The best score per language and criterion is shown in bold.

genre and author (Prompt 2, 3a) or a request of creativity (Prompt 3, 3b).

For all our languages, Prompt 3 (3b) has better solutions than Prompt 2 (3a) as it generates more CSs and fewer errors. When compared to the results of the other phases, we also see that Prompt 3 has the best performance overall for ES, NL and ZH, with the most number of CSs and the least number of errors. However, for CA, the best performance was in Phase 1 (1b), with Prompt 1 at the document level. The explanation for this lies again in the translation of sobriquets, which are left untranslated in both 3a and 3b. In fact, the only settings in which sobriquets are translated at all are paragraph level (6 out of 14, as said above) and document level (all of them). Since 4 sobriquets are UCPs classified independently, their translations impact the formula. If the sobriquets were excluded, 3a and 3b would be the best-performing settings for CA. The sobriquets were also problematic for ZH and ES: for ES, 3a kept all sobriquets in English and 3b did not translate 2 out of 14 sobriquets; for ZH, it was 3b that did not translate the words but kept them in English, although 3b output had better performance than 3a or any of the other outputs. Surprisingly, for NL, both 3a and 3b translated the sobriquets into Dutch, although 3a retained one sobriquet in English. This might explain the relative high score for NL with 3b compared to the other languages.

4.4 ChatGPT vs Others

We also compare the performance of ChatGPT with that of DeepL and since CA is not available in the latter, we use GT (ENCA-G) and Softcatalà (ENCA-S). At the time we ran DeepL its new-gen version was available for ZH but not for ES nor NL. Therefore we used DeepL new-gen for ZH and DeepL classic for ES and NL. The creativity index of these baseline systems are shown in Table 5.

	# CSs	# Errors	Error Points	CI
ENCA-S	1	83	393	-63.43
ENCA-G	1	66	261	-41.50
ENES	9	58	237	-22.70
ENNL	6	51	103	-6.00
ENZH	11	42	152	-4.88

Table 5: Creativity Index in Others (3c). S stands for Softcatalà's *Traductor* and G for Google Translate.

In all languages, the selected NMT system (3c) performs worse than the best setting of ChatGPT. In ZH, NL and ES, DeepL performs better than some of the other settings in ChatGPT, while in CA the two NMT systems perform worse than all ChatGPT outputs. This shows that ChatGPT, with an appropriate prompting strategy, has the potential to outperform its NMT counterparts in literary text in terms of creativity.

5 Analysis

We wanted to further analyse the MT performance in all the phases, and also compare the best performing setting with the professional translation

described in Section 3.6. Firstly, Figure 2 and Figure 3 show the comparison between ChatGPT in the different settings in terms of CSs and Error points (including Others as in Section 4.4).

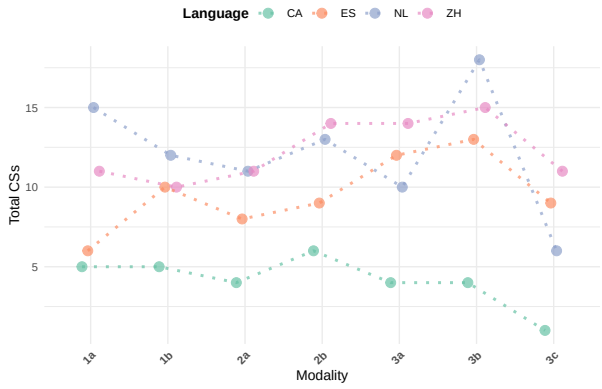


Figure 2: Total CSs per Modality and Language

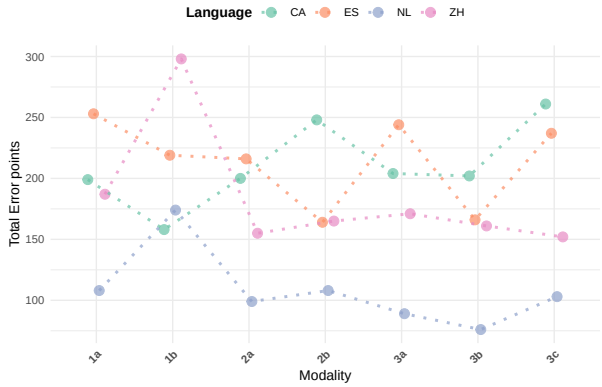


Figure 3: Total Error points per Modality and Language

Figure 2 and 3 illustrate the results already illustrated in Tables 2, 3, 4, 5 more clearly. To assess the effect of Modality and Language on CSs, an Aligned Rank Transform (ART) ANOVA was conducted for non-parametric data. Results show a significant main effect of Language, F(3, 1431) = 30.26, p = .000. However, there are no effects of Modality or the interaction of Modality and Language. Pairwise comparisons using Bonferroni correction show that CSs was significantly lower in CA than ES, NL and ZH p = .000. This is somewhat logical as the number of CSs is very low in all settings, and even lower in CA. We then assess the effect of Modality and Language on Error Points, the results show a significant effect of Modality, F(6, 1431) = 4.63, p = .000, and Modality × Language interaction, F(18, 1431) = 1.7, p = .03. The pairwise comparisons show that Error points was significantly higher in 1b when compared to 2a, p = .000, and to 3b, p = .001, and 2b was significantly higher than 3b, p = .025. This shows again that 1b and 3b were the best performing settings for these languages. The interaction analysis shows only a significant result between 3b/NL and 3c/CA.

Secondly, Figures 4 and 5 illustrate the comparison of the best performing setting with the professional translations in terms of CSs and Error points. To assess the effect of Modality and Language on CSs, we created a subset by grouping the best performing setting under the variable MT to compare it to HT. The ANOVA indicates a significant main effect of Modality (only HT and MT in this case), F(1, 265) = 31.70, p = .000, and Language, F(2, 265) = 4.26, p = .015. There was no effect of the interaction of Modality and Language. A pairwise comparison shows that CS was significantly higher in HT than in MT (p = .00). The effect of Language was only significant for CA and NL (p = .02), but not for CA and ES or ES and NL.

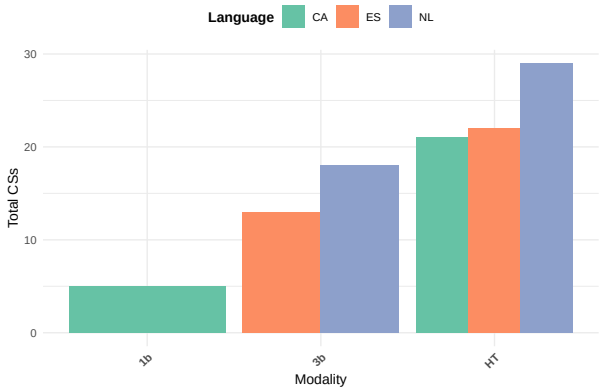


Figure 4: Total CSs per best ChatGPT Modality and HT

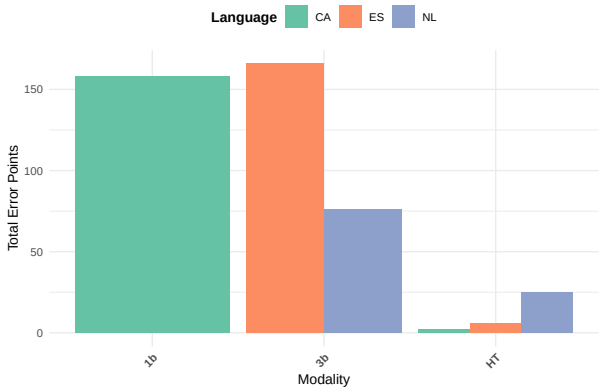


Figure 5: Total Error points per best ChatGPT Modality and HT

For Error points the results show a significant main effect of Modality, F(1, 265) = 46.27, p = .00,

Language, F(2, 265) = 5.21, p = .006, and Modality × Language interaction, F(2, 265) = 10.66, p = .00. A pairwise comparison shows that Error points was significantly lower in HT than in MT (p = .00). The effect of Language was only significant for ES and NL (p = .005), but not for CA and ES or CA and NL. When looking at the interactions, the comparison of HT/Languages and MT/Languages, there is significance in all the combinations of HT and MT (p = .00) except in the interaction between HT/NL and MT/NL.

5.1 Analysis of AEMs

Our main interest in running a set of representative AEMs (see Section 3.5), is to find out whether any of them correlates significantly with any of the metrics used in the human annotation (i.e. CSs, error points, CI). A limitation in this regard is that the number of instances[15] is very small, which is why we use a non-parametric correlation metric (Spearman). We find, as expected, significant correlations between pairs of AEMs, which occur most often between pairs of string-based metrics. Only for one language (NL) do we find significant correlations between one human metric (CSs), and two AEMs: chrF ($p < .001$) and COMET-Kiwi ($p < 0.05$). Given, again, the small sample size, and that they occur only in two cases, we refrain from drawing any strong conclusion.

We also calculated detailed scores for the TER metric. Namely, the number of operations per operation type (insertions, deletions, shifts and substitutions), system and language. The main observation is that across all languages and systems, the number of substitutions (range $[960, 1286]$) is considerably higher than the number of the other operation types put together: insertions ($[149, 290]$), deletions ($[102, 225]$) and shifts ($[95, 129]$). All the scores with AEMs are reported in Appendix B.

6 Conclusions

We wanted to explore ChatGPT MT for the best possible setting for creativity. The results show that there is indeed variability per configuration and per language. The first observation, perhaps obvious for a translator but not so obvious for others, is that creativity is seriously affected by using ChatGPT in any setting. Not only is the number of CSs in the TTs provided by all ChatGPT models (but also DeepL, GT and Softcatalà) significantly lower than in HT, but the number of errors is also significantly higher. Even the most creative setting does not come close in three out of the four languages analysed (for ZH we did not have an HT reference). Further, it is important to note that the CI for HT is not only higher but it might also not be representative of the overall creativity of the HT TTs, since we are only analysing the solutions provided by the translators to the annotated UCPs but not the entire segment where translators use other techniques, e.g. compensations, to create the desired overall effect of the text.

The second observation is that less appears to be more when prompting ChatGPT to output a creative translation. Overall the best result is the one provided by **Prompt 3**: "Translate the following text into [TG] creatively". Although this prompt still yields a very high number of errors and very modest CSs, it still outperforms the others in ES, NL and ZH while in CA even less information is needed as **Prompt 1**: "Translate the following text into [TG]" outperforms the others. These results are in line with the previous results obtained for Chinese in Du (2024).

The different prompts have somewhat similar results across different languages, with better outputs for temperature 1.0 (2b) and with Prompt 3 (3b) for ES, NL and ZH, although there were differences when providing ChatGPT with paragraphs or the whole document and between CA and the other languages. Moreover, it is interesting to see that there is a level of randomization in the output that is quite unpredictable and that requires many iterations to find the optimal solution. We wonder how this fits in a context where MT is supposed to be used to increase translator's performance. Trying these different alternatives and still obtaining a sub-optimal result does not seem the best solution for practicing translators, although it is impossible to predict if some MT suggestions might spark creativity.

As this case study is of an exploratory nature, there are limitations, notably, we selected a reduced number of UCPs that were annotated by one single annotator, with a limited number of prompts. However, the striking differences in the performance in literary translation in comparison to what is reported in the media, i.e. singularity (Translated, 2025), merits urgent attention.

[15] i.e. number of modalities per language: $n = 8$ for CA and $n = 7$ for the other three languages.

Acknowledgments

This project has received funding from the EU ERC Consolidator Grant 101086819; a Beatriz Galindo senior fellowship (BG23/00152) from the Spanish Ministry of Science and Innovation; and Grant PID2023-150711OB-I00 funded by MICIU/AEI/10.13039/501100011033 and ERDF/EU.

Sustainability statement

All in all, we submitted 2,586 API requests to ChatGPT, leading to the processing of 436,054 tokens (combining inputs and outputs). To the best of our knowledge, the average CO^2 emissions of GhatGPT models is not disclosed. A calculation by a third party estimates that each message sent to ChatGPT produces approximately 4.32g CO^2 (Wong, 2024). Asking ChatGPT we obtain the range $[2.5, 23.75]$, depending on the electricity source and assuming 50 Wh per query. Using the 4.32g CO^2 figure above, our experiments would have emitted 11.2kg CO^2.

It is also worth taking into account that we submit two rather different types of queries: paragraph- and document-based. For a paragraph-based translation we submit 125 queries, which take around 2 minutes and 50 seconds, i.e. 1.36 seconds per query. For a document-based translation only 1 query is sent, which takes around 2 minutes and 7 seconds.

References

Gerrit Bayer-Hohenwarter. 2009. *Translational creativity: how to measure the unmeasurable*, volume 37. Samfundslitteratur Copenhagen.

Gerrit Bayer-Hohenwarter. 2011. "Creative Shifts" as a Means of Measuring and Promoting Translational Creativity. *Meta Journal des traducteurs*, 56(3):663–692.

Shuxiang Du. 2024. *Optimizing Creative Translations through ChatGPT: An analysis of the Creative Potential of Machine Translation in Literary TextsCommunication and Information Studies*. Master's thesis, University of Groningen.

Margot Fonteyne, Arda Tezcan, and Lieve Macken. 2020. Literary machine translation under the magnifying glass: Assessing the quality of an nmt-translated detective novel on document level. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3790–3798.

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*, pages 1–7.

Ana Guerberof-Arenas and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, 9(2):255–282.

Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in translation. *Translation Spaces*, 11(2):184–212.

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*.

Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762.

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. *arXiv preprint arXiv:2301.08745*.

Katikapalli Subramanyam Kalyan. 2023. A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 6:100048.

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.

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.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica tecnologies de la traducció*, (12):455–463.

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 on Association for Computational Linguistics, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*.

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.

Ziqian Peng, Rachel Bawden, and François Yvon. 2024. Investigating length issues in document-level machine translation. *arXiv preprint arXiv:2412.17592*.

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.

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. 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.

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.

Arda Tezcan, Véronique Hoste, and Lieve Macken. 2017. Scate taxonomy and corpus of machine translation errors. *Trends in E-tools and resources for translators and interpreters*, 45:219–244.

Translated. 2025. Discover How Close We Are to AI Singularity.

Kurt Vonnegut. 1999. *Bagombo Snuff Box*. Putnam Adult.

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.

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.

Vinnie Wong. 2024. Gen AI's Environmental Ledger: A Closer Look at the Carbon Footprint of ChatGPT. https://piktochart.com/blog/carbon-footprint-of-chatgpt/. Accessed: 2025/02/07.

A Detailed Human Annotations

Tables 6, 7, 8 and 9 show the detailed human annotations of all languages. The best condition per language is shown in bold.

ENZH	Para	Doc	T-0.0	T-1.0	Prompt2	Prompt3	DeepL
Abstraction	1	1	1	**2**	0	0	1
Concretization	5	5	5	5	4	**8**	4
Modification	5	4	5	7	**10**	7	6
Reproduction	35	**30**	39	35	35	32	36
Omission	4	5	4	2	2	**1**	**1**
Error in UCPs	4	9	**0**	3	3	6	6
#CSs	11	10	11	14	14	**15**	11
#Errors	51	56	45	51	39	**37**	42
Error Points	187	298	155	165	171	161	**152**
Score	-10.69	-30.98	-5.38	-1.48	-2.48	**1.03**	-4.88

Table 6: Detailed human annotation - ENZH

ENNL	Para	Doc	T-0.0	T-1.0	Prompt2	Prompt3	DeepL
Abstraction	2	**5**	3	4	3	2	1
Concretization	**4**	3	2	2	3	**4**	2
Modification	9	4	6	7	4	**12**	3
Reproduction	**33**	37	40	38	42	**33**	44
Omission	0	0	0	0	0	0	1
Error in UCPs	**2**	5	3	3	**2**	3	3
#CSs	15	12	11	13	10	**18**	6
#Errors	61	68	49	51	43	**30**	51
Error Points	108	174	99	108	89	**76**	103
Score	9.84	-6.68	3.93	6.13	3.73	**20.71**	-6.00

Table 7: Detailed human annotation - ENNL

ENES	Para	Doc	T-0.0	T-1.0	Prompt2	Prompt3	DeepL
Abstraction	1	**2**	1	1	**2**	1	1
Concretization	2	2	3	2	**4**	**4**	1
Modification	3	6	4	6	6	**8**	7
Reproduction	42	43	44	42	40	**38**	**38**
Omission	1	**0**	1	1	**0**	**0**	2
Error in UCPs	5	**1**	**1**	2	2	3	5
#CSs	6	10	8	9	12	**13**	9
#Errors	59	52	55	50	57	**43**	58
Error Points	253	219	216	**164**	244	166	237
Score	-31.00	-18.00	-21.00	-11.00	-18.31	**-3.50**	-22.70

Table 8: Detailed human annotation - ENES

ENCA	Para	Doc	T-0.0	T-1.0	Prompt2	Prompt3	Softcatalà	Google Translate
Abstraction	0	0	0	0	0	0	0	**1**
Concretization	**1**	**1**	0	**1**	0	0	0	0
Modification	4	4	4	**5**	4	4	1	0
Reproduction	42	46	47	45	47	47	**39**	45
Omission	1	1	1	1	1	1	1	1
Error in UCPs	6	2	2	2	2	2	13	7
#CSs	5	5	4	**6**	4	4	1	1
#Errors	**51**	**51**	57	69	57	55	83	66
Error Points	199	**158**	200	248	204	202	393	261
Score	-23.80	**-16.99**	-25.82	-30.08	-26.48	-26.15	-63.43	-41.50

Table 9: Detailed human annotation - ENCA

System	**BLEU**			**chrF**			**TER**			**COMET**			**COMET-Kiwi**			
	ENCA	ENES	ENNL	ENCA	ENES	ENNL	ENCA	ENES	ENNL	ENCA	ENES	ENNL	ENCA	ENES	ENNL	ENZH
1a	24.8	23.0	28.9	51.9	51.1	55.3	61.4	65.0	55.8	0.781	0.7641	0.8254	0.7857	0.8033	0.8223	0.8075
1b	23.1	22.3	26.3	50.4	50.3	53.2	63.0	64.7	58.1	0.7687	0.7482	0.8101	0.7804	0.7918	0.8037	0.6098
2a	25.6	23.2	29.5	52.0	51.0	56.3	60.6	64.5	55.2	0.7744	0.7567	0.8281	0.7935	0.8000	0.8252	0.8089
2b	23.4	22.0	29.0	50.8	50.4	55.9	63.0	65.2	55.7	0.7693	0.7650	0.8248	0.7753	0.8049	0.8252	0.8093
3a	26.0	22.5	28.6	52.3	49.9	55.5	60.6	65.8	56.8	0.7736	0.7569	0.8253	0.7908	0.7983	0.8276	0.8092
3b	25.4	22.7	25.0	51.7	50.2	53.8	61.3	65.2	61.6	0.7703	0.7604	0.8238	0.7880	0.7955	0.8163	0.8017
3c	21.7	25.1	31.8	47.9	51.7	55.9	65.4	62.5	54.3	0.7158	0.7714	0.8257	0.7495	0.8078	0.8301	0.8077
3d	25.3			51.3			61.2			0.7576						0.7714

Table 10: Scores with a set of AEMs for each system and language pair.

B Scores with Automatic Evaluation Metrics

Table 10 shows the scores for each system and target language with a set of representative automatic evaluation metrics (see Section 3.5), while Figure 6, Figure 7 and Figure 8, show TER's number of operations per operation type for each system and target language.

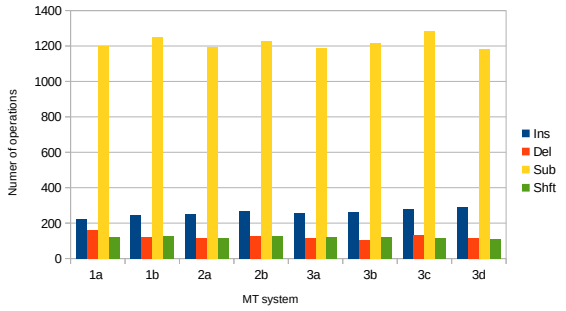


Figure 6: TER's number of operations per operation type (insertions, deletions, substitutions and shifts) for English⇒Catalan

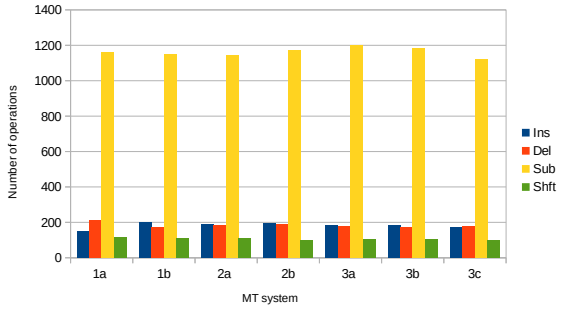


Figure 7: TER's number of operations per operation type (insertions, deletions, substitutions and shifts) for English⇒Spanish

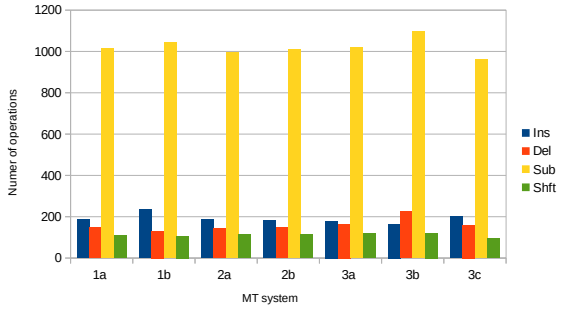


Figure 8: TER's number of operations per operation type (insertions, deletions, substitutions and shifts) for English⇒Dutch

Improving MT-enabled Triage Performance with Multiple MT Outputs

Marianna J. Martindale† and **Marine Carpuat**‡
†College of Information, ‡Department of Computer Science
University of Maryland, College Park, USA

Abstract

Recent advances in Machine Translation (MT) quality may motivate adoption in a variety of use cases, but the success of MT deployment depends not only on intrinsic model quality but on how well the model, as deployed, helps users meet the objectives of their use case. This work focuses on a specific triage use case, MT-enabled scanning in intelligence analysis. After describing the use case with its objectives and failure modes, we present a user study to establish a baseline performance level and measure the mitigating effects of a simple intervention, providing additional MT outputs. We find significant improvements in relevance judgment accuracy with outputs from two distinct neural MT models and significant improvements in relevant entity identification with the addition of a rule-based MT. Users also like seeing multiple MT outputs, making it an appealing way to improve MT-enabled scanning performance.

1 Introduction

Recent years have seen dramatic advances in Machine Translation (MT) quality (Kocmi et al., 2022, 2023, 2024), making MT adoption in a variety of use cases all the more appealing. But intrinsic model quality does not dictate success or failure in MT deployment. For any given use case, the critical question is not how well the model performs on benchmark evaluations, but how effectively the model, as deployed, will help users accomplish their objectives. That requires understanding the objectives of the use case as well as the strengths and weaknesses of MT.

In this work, we focus on a triage use case, MT-enabled scanning in intelligence analysis, and its objectives and failure modes (Section 2.1). We will then discuss how the strengths and weaknesses of available MT systems may affect user performance (Section 2.2) and interventions that might improve performance (Section 2.3). Finally, we will detail our user study (Section 3) and provide recommendations for this and similar use cases based on the results (Section 4).

2 Background

2.1 MT-Enabled Scanning Use Case

In this work we refer to the process in intelligence analysis of labeling documents as relevant (to be kept for further analysis) or NTR (Nothing To Report) as “scanning". Like many triage use cases, scanning involves volumes of text large enough that it is impractical to have people who know the language perform triage. Instead, users familiar with the domain who do not know the language use MT to identify documents believed to be relevant enough to send for human translation. Because the users don’t know the language, they are susceptible to misleading errors in the MT output, but the risk of incorrect information from the MT output ending up in intelligence reports is mitigated by human translation before further analysis. However, MT errors that mislead the user still incur costs from irrelevant documents, wasting human translator time, or bear a risk of missing relevant documents.

2.2 Reliability of MT

Although there are no prior studies on MT-enabled triage for intelligence analysis, prior work on the reliability of MT can help us understand how the strengths and weaknesses of MT may affect this use case. Older MT approaches, such as statistical and rule-based MT (RBMT), suffered from fluency issues that can lead users to distrust the output (Martindale and Carpuat, 2018). The improved fluency of generated output comes with an increased risk of output that is detached from the meaning of the input, often referred to as hallucinations. This trend

Proceedings of Machine Translation Summit XX Volume 1, pages 592–607, June 23-27, 2025

was initially observed in the earliest Neural MT (NMT) models (Koehn and Knowles, 2017; Lee et al., 2018; Martindale et al., 2019; Raunak et al., 2021) but has remained an issue in more recent MT models (Xu et al., 2023; Guerreiro et al., 2023) and Large Language Models (LLMs) (Kalai and Vempala, 2024). Despite their fluency, hallucinations may not be believable in context (Martindale et al., 2021), but if believable in context, the user will be misled. Without intervention, the user must rely on surface features such as fluency, document context, and real-world context in deciding whether the MT output is an accurate representation of the meaning of the source text.

2.3 Possible Interventions

There are many possible interventions that could reduce how often users are misled during MT-enabled triage tasks. Our interventions should help users calibrate their judgments of MT output to decrease the believability of errors while increasing the believability of accurate translations. Explainability approaches such as confidence scores may help users calibrate trust in AI models (Zhang et al., 2020), but users may still be misled by low-confidence incorrect output (Suresh et al., 2020) and can have difficulty detecting critical errors (Mehandru et al., 2023). For MT in particular, sentence-level confidence scores tend not to be well-calibrated without explicitly adapting the training to encourage better calibration (Kumar and Sarawagi, 2019; Wang et al., 2020; Lu et al., 2022) and lack the specificity needed to help the user decide which parts of the translation to believe. Fine-grained MT quality estimation (QE) approaches like those in WMT shared tasks on word-level QE (Specia et al., 2021) and fine-grained error span detection (Blain et al., 2023) provide additional information for the user, but the best models do not perform well enough and require considerable resources, with the top submissions in WMT23 only achieving F1 scores below 0.3 for models with as many as 13B parameters or ensembles of up to 12 models (Blain et al., 2023). Rather than simply highlighting error spans in the output, Briakou et al. (2023) improve explainability using contrastive phrasal highlights to draw the reader's attention to meaning differences. The approach was tested with bilingual users in a human translation quality review scenario, but monolingual users could apply linguistic resources such as dictionaries to the highlighted source text phrases to verify the severity of divergence. This is a promising approach, but it is unclear whether the current models are performant enough for deployment without significant engineering effort.

The ideal intervention can immediately be deployed with MT models of any quality and will have the potential to continue to help users even as newer, better models are deployed. The best fine-grained error detection model at WMT23 relied on pseudo-reference translations generated by off-the-shelf MT systems (Rei et al., 2023). What if we simply provided the user with the alternate translation? This type of intervention is appealing because it requires no additional data or specialized skills and can be used for any language where more than one MT system is available. Prior work has shown that displaying two MT outputs improves confidence and performance in MT-mediated communication without increasing cognitive load (Xu et al., 2014; Gao et al., 2015). We hypothesize that MT-enabled triage use cases can derive similar benefits.

3 User Study Design

To establish a baseline risk level for MT-enabled triage in intelligence analysis and to measure the mitigating effects of practical interventions, we conducted a user study with Intelligence Analysts (IAs) from a US intelligence agency in the Washington, DC area with significant experience (at least three months) performing triage tasks with the aid of MT and little or no knowledge of the source language. In the next sections, we describe our interventions and design a scenario and tasks for the user study that mimic real triage tasks. We then address the format of the user study and analysis methods.

3.1 Intervention: Multiple MT Outputs

To mitigate the risks of misleading MT output, we propose two versions of the alternative translations intervention from Section 2.3 (pairing output from a single NMT system with output from a second NMT[1] system, and pairing a single NMT output with rule-based MT (RBMT) output). We also propose a combination of the two versions, displaying two NMT outputs with RBMT output.

IAs with output from only one MT system must rely on features of the output text, like fluency, and

[1]Note: LLMs were not yet available when the data for the user study tasks was translated and annotated. See Section 6

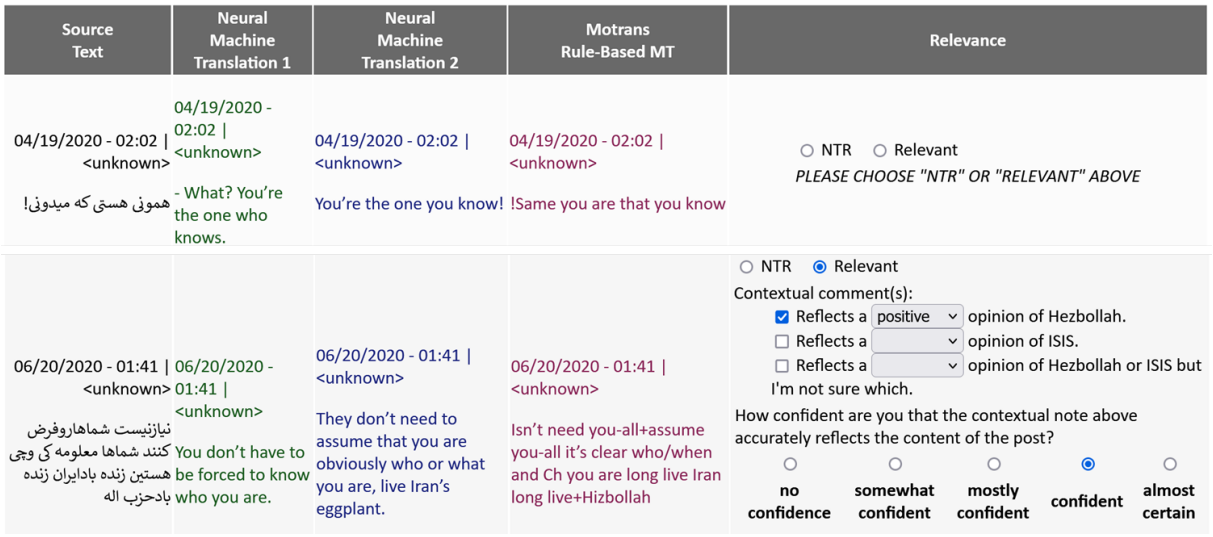


Figure 1: The user interface for a Hezbollah/ISIS conversation thread.

contextual features, like plausibility, to decide the extent to which they believe an MT output reflects the meaning of the source text. A second MT output provides additional information to inform the decision. Differences between the two translations will draw attention to potential errors in fluent output and similarities between the translations can overcome disfluencies that would otherwise reduce the believability of an MT output.

In the second version of the intervention, RBMT output is not expected to provide the readability of neural MT output but does provide more interpretability than off-the-shelf NMT because every word or phrase in the output is a translation of specific words in the source. It is also easy to update with new named entities and specialized terminology, making it especially useful for keyword-spotting. For these reasons, the CyberTrans MT platform (Reeder, 2000) available to analysts throughout the US Intelligence Community includes Motrans RBMT for many languages (Martindale, 2012). Paired with one NMT output, Motrans can provide a similar effect to displaying a second NMT output if the output is sufficiently readable or contains relevant keywords. Paired with two NMT outputs with significant meaning differences, the Motrans output's reliable connection to the source can make it a useful "tie-breaker".

3.2 User Study Tasks

This study focuses on Persian Farsi conversation threads in a scenario intended to be analogous to real intelligence analysis use cases. Persian was selected as the language for the study because it is of strategic importance and poses challenges for MT due to the limitations of available training data but there are open-source pre-trained models and commercial-off-the-shelf software available that can translate from Persian to English, as well as a Motrans capability. Conversation threads were chosen as our documents because, due to their difficulty, performance on conversation threads may be seen as a lower bound on analyst relevance judgment performance more broadly. Understanding any given message requires understanding its context in the conversation, and conversational text also often uses colloquial language which may be out of domain for MT systems.

For reasons of security and practicality, it is not possible to conduct the study using conversation threads from analysts' actual data, so this study relies on an analogous collection of publicly available data gathered from user comments on Persian-language news articles. The topics for the user study are: Opinions related to the Russia-Ukraine conflict and Opinions related to terrorist organizations, specifically Hezbollah and ISIS. These topics were chosen because they relate to US intelligence priorities (strategic competition and violent extremist organizations) and are likely to elicit reactions among readers of Iranian news articles because of Iran's support of Russia (Bowen et al., 2022) and Hezbollah (Humud, 2023) and Iran's stance against ISIS (Arango and Erdbrink, 2014).

Analogous to the real MT-enabled triage use case, participants were asked to identify high-level features based on MT output in context. Each task consisted of one or more conversation threads that

the user must scan for comments relevant to key intelligence questions, which they would label as either "Relevant" or "NTR" (Nothing to Report). They were also asked to identify information in the relevant comments as if they were adding a context note when passing the document to be translated. Finally, they were asked to rate their confidence in their judgments. A screenshot of the user interface for a Hezbollah/ISIS task conversation thread with both NMT outputs and Motrans RBMT output is shown in Figure 1, with the first comment unannotated and the second comment displaying the contextual note options.

The contextual note information was gathered in a multiple-choice, fill-in-the-blank style. Analysts could choose whether the comment is related to one or both of the relevant entities and whether the comment expresses a positive or negative opinion of that entity. Analysts could express uncertainty about the target of the comment by choosing an option that says they believe the comment reflects an opinion of one of the entities but they are not sure which. They could also express uncertainty about the stance of the comment by choosing "unclear" rather than "positive" or "negative." This allows for a granular evaluation of comprehension, from relevance judgment to information extraction to stance detection.

During the post-task survey, participants provided feedback validating the similarity of the tasks to their typical work, as discussed in Section 4.1.

3.3 Data and Annotation

The initial corpus of comment threads was collected in July 2022 by searching Persian-language news sites[2] for Farsi keywords related to the topics and then scraping the user comments, replies, and their publicly visible metadata (username, timestamp, and threading information) from the articles that were returned. Filtering for threads with at least two replies yielded 1,552 comments in 315 threads for Russia-Ukraine and 346 comments in 82 threads for the terrorism topic. Given limited annotation resources, we further filtered the Russia-Ukraine comments by selecting threads that were more likely to contain at least one comment with a potentially misleading translation in the context of this task using the Twitter-trained sentiment analysis model from TimeLMs (Loureiro et al., 2022) on the MTs of the comments and choosing threads that contained at least one comment for which the two NMT outputs had different sentiment labels. This resulted in 210 comments in 35 threads for annotation from the Russia-Ukraine topic.

The MT systems for the user study were chosen based on fitness for the use case. Because hallucinations are often tied to the training data (Raunak et al., 2021) we expect that output from a second model trained on different data is unlikely to produce the same hallucinations, so we want our NMT models to have been trained on substantially different data. One way to know the models were trained on different data is to use a bilingual model and a multilingual model, ensuring that even if both models were trained on similar Persian-English bitext, the multilingual model will have been exposed to additional English target text for other language pairs. To this end, we use a freely available massively multilingual pre-trained model, NLLB-200 (Koishekenov et al., 2022), and a commercial off-the-shelf system, SYSTRAN (version SPNS 9.7) as our two NMT systems. Open-source pre-trained models like NLLB-200 are appealing because they can be deployed on an intranet with minimal machine learning knowledge, and NLLB-200 is particularly desirable because it covers 200 languages, making it a logical choice for our baseline NMT system. SYSTRAN is a plausible second NMT system because it is familiar to US government users through long-standing collaboration with the Air Force (SYSTRAN, 2021) and previous integration in government translation platforms such as CyberTrans (Reeder, 2000).

The Persian Motrans capability that produced our RBMT outputs was developed from electronic dictionaries in the mid-2000s and continues to be updated with technical terms, named entities, and colloquialisms observed in sources such as news, technical documents, and web content. Motrans is optimized for adequacy rather than fluency. It handles ambiguity by providing alternative translations separated by a slash in the output and it attempts to split out of vocabulary tokens into smaller translatable words with '+' between the resulting translations in the output, as shown in the Motrans translation of the second comment in Figure 1. The ambiguous Farsi word کی is translated as *who/when*, and the incorrectly spaced phrase زنده بادحزب اله is translated as *long live+Hizbollah*.

Two Persian language analysts were recruited to provide gold standard annotations on the comments. The annotators completed the same relevance judg-

[2]isna.ir/news, tabnak.ir/fa/news, and khabaronline.ir/news

ment and contextual note task that user study participants would complete but using the source text rather than the MT output. They also evaluated the MT output quality using a task-focused adaptation of the evaluation scales from Licht et al. (2022). Each quality level was given a descriptive label to emphasize that they are labels rather than equally distanced points in a range. The lowest quality label was *MISS*, described as a translation that is so different from the meaning of the source text that a non-language-enabled analyst would not be able to reliably make even a relevance judgment. The second level was *REL-ONLY*, described as translation quality sufficient to make a relevance judgment but with significant information missing or incorrect. The third level was *GIST*, described as translating critical information correctly but with less important information missing or incorrect. Levels 4 and 5 were labeled *GOOD* and *EXCELLENT* respectively. Translations below GIST quality (MISS or REL-ONLY) can be considered potentially misleading in this scenario. Details can be found in Appendix A.

From the annotated comments, we selected conversation threads to use in two tasks per topic, each totaling approximately 15 comments. The threads were selected based on the relevance and MT quality judgments with the goal of including at least one unambiguously relevant comment per thread, at least two comments in each task where NMT1 was potentially misleading and NMT2 was GIST or better, and at least two comments where NMT1 was potentially misleading and Motrans was GIST or better. Of the 61 comments included in the user study, 32 were labeled relevant.

3.4 User Study Methods

The user study was conducted with a 2x2 design, with one between-subjects variable and one within-subjects variable. The between-subjects variable is whether the analyst sees one NMT output or two and the within-subjects variable is whether the analyst is provided rule-based MT output from Motrans in addition to the NMT output(s). Participating analysts were assigned to either the one-NMT or two-NMT condition and completed tasks both with and without Motrans output provided. The order of presentation of the conditions (with and without Motrans) was counterbalanced across analysts to control for ordering effects. Each analyst completed two tasks in each condition, and the order of all four tasks was counterbalanced to control for task-specific ordering effects.

The study was reviewed and approved by the University of Maryland Institutional Review Board, protocol number 1964637-1, and the Human Research Protection Program of the agency where the study took place. Informed consent was obtained from all participants prior to data collection.

Participants were recruited through messages on internal networking sites and mailing lists, with the goal of recruiting up to 40 qualified IAs. They were screened using a qualification survey, which also gathered relevant background information for qualified participants and asked about their perceptions of the MT they use. When they completed the background survey, their responses were validated against participation criteria, and if they qualified, they were asked to commit to completing the user study on a specific date and time of their choosing. Those who provided a date and time were assigned a batch of tasks round-robin style. In total, 35 IAs responded to the survey, and 26 completed the user study. Two of the survey respondents did not qualify because their MT use was in a language they knew and seven analysts did not respond to contact after the background survey. Two of the remaining 28 IAs, both from the 1-NMT condition, failed to complete the user study as assigned, leaving 12 participants in the 1-NMT condition and 14 in the 2-NMT condition.

4 Results

The discussion of the user study results is structured as follows. First, we validate the user study scenario and tasks. We then establish the baseline performance using output from one NMT system and demonstrate the mitigating effects on performance from providing additional outputs, followed by the effects on confidence. After summarizing these quantitative results, we briefly address acceptability of the interventions as indicated by responses on the pre- and post-task surveys.

4.1 Scenario Validation

Responses in the post-task survey verified whether the user study tasks were similar to intelligence analysis foreign language triage tasks. No participants said the tasks were “Much Easier" or “Much Harder" while 15% of users said that they were easier than their “typical foreign language text triage tasks," 42% said they were of similar difficulty, and 42% said they were harder. In open ended-

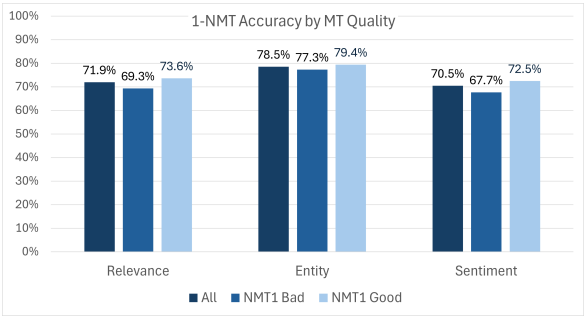


Figure 2: Mean accuracy in the 1-NMT condition across all examples (All) and with NMT1 quality below GIST (NMT1 Bad) and GIST or better (NMT1 Good), showing highest accuracy for entities and lowest for sentiment with a 4.8 point difference in sentiment accuracy between NMT1 Good and Bad.

responses regarding the elements of the user study that were similar to or different from analysts' typical triage tasks, the most frequently mentioned similarity was the overall framing, mentioned by 11 analysts. Five analysts mentioned similar MT quality and four analysts mentioned similar task difficulty. Three analysts noted that similar to their tasks, the conversations lacked context and the comments were short and informal. However, two analysts cited the length of the text as a difference, noting that they typically triage whole documents. Other differences that were mentioned included topic (seven analysts), language (three analysts), and their familiarity with the topic (five analysts). Only one analyst mentioned a difference in the structure of the task, stating that they do not typically write contextual notes "but its [sic] a good idea." Overall, these responses indicate that the scenario for the user study is comparable to many analyst workflows and the conversation threads selected are analogous to at least some real-world MT-enabled triage use cases.

4.2 Baseline Analyst Performance

Relying on output from only one NMT system, users (n=12) averaged 70% or higher accuracy on all three levels of comprehension, as seen in Figure 2. The entity accuracy score is highest (nearly 80%), likely because it is often possible to quickly tell when a comment refers to an entity by spotting the entity's name. The sentiment score is lowest (70.5%), supporting the intuition that identifying the stance towards the subject of a comment is more difficult than just identifying the subject of the comment.

Partitioning the comments based on the quality of the output from NMT1, we can measure performance on the potentially misleading examples (*NMT1 Bad*) compared to the *NMT1 Good* examples. For all three accuracy measures (relevance, entity, and sentiment), we see that mean accuracy is lower when the MT quality is bad (below GIST) and higher when the quality is good (GIST or better). The biggest difference is in sentiment, where the mean accuracy for bad translations is 67.7% compared to 72.5% for good translations.

The overall baseline accuracy reflects the utility of the baseline NMT system for this triage task, but leaves significant room for improvement, even when the MT output is fairly high.

4.3 Impact of Interventions on Accuracy

As described in Section 3.4, the user's response to each comment provides three labels we can score for accuracy: relevance judgment, sentiment for Entity A, and sentiment for Entity B. Each user's responses were compared against the gold standard annotations. We want to see whether adding a second NMT, adding RBMT, or the combination of both significantly affects relevance, entity, or sentiment accuracy, so we build three Generalized Linear Mixed Effects Models (GLMM) (one with each type of accuracy as the response variable) with fixed effects for the presence of a second NMT, presence of RBMT, and interaction between NMT and RBMT. We used random effects to control for user and item. For each GLMM, there are 1586 observations, grouped by item (61) and user (26).

Results of the GLMM with Relevance Accuracy as the response variable are shown in Table 1. We see a significant ($p < 0.05$) increase from adding a second NMT (OR=2.26, CI=1.35-3.85, p=0.0032) as well as adding RBMT (OR=1.52, CI=1.37-3.74, p=0.041) and a significant interaction from providing both (OR=0.52, CI=0.29-0.91, p=0.03). Based on this odds ratio, a hypothetical analyst with 3:1 odds of being correct in their relevance judgments with only one NMT would have their odds increased to 6.8:1 with a second NMT output. The same analyst would have their odds increased to 4.5:1 with the addition of RBMT. Note the negative β value for the interaction between NMT and RBMT. This means that although we would expect adding both a second NMT and RBMT to increase the analyst's odds of being correct to 10.2:1, the interaction effect means the odds only increase to 5.3:1, which is higher than just adding RBMT but

Coefficient	β	Odds Ratio	Confidence Interval	p
(Intercept)	1.497	4.469	2.476 - 8.067	< 0.001
2-NMT	0.816	2.262	1.370 - 3.735	0.0014
w/ RBMT	0.416	1.516	1.018 - 2.259	0.0408
2-NMT+RBMT	-0.660	0.517	0.294 - 0.909	0.0219

Table 1: GLMM for Relevance Accuracy showing largest significant (p<0.05) effect from the second NMT.

Coefficient	β	$exp(\beta)$	Confidence Interval	p
2-NMT	0.479	1.614	0.719 - 3.627	0.3627
w/ RBMT	0.554	1.740	1.204 - 2.514	0.0032
2-NMT+RBMT	-0.505	0.603	0.362 - 1.007	0.0530

Table 2: CLMM for Entity Accuracy showing significant improvement (p<0.05) from adding RBMT.

Coefficient	β	$exp(\beta)$	Confidence Interval	p
2-NMT	0.456	1.578	0.433 - 1.485	0.1540
w/ RBMT	0.251	1.285	0.933 - 1.770	0.1250
2-NMT+RBMT	-0.331	0.718	0.460 - 1.122	0.1460

Table 3: CLMM for Sentiment Accuracy showing no significant effects.

Coefficient	β	$exp(\beta)$	Confidence Interval	p
2-NMT	1.052	2.863	1.098 - 7.466	0.0315
w/ RBMT	0.037	1.038	0.780 - 1.382	0.7980
2-NMT+RBMT	0.098	1.103	0.755 - 1.612	0.6130
Relevance Accuracy	0.349	1.417	0.924 - 2.174	0.1100
Entity Accuracy	-0.404	0.667	0.264 - 1.691	0.3940
Sentiment Accuracy	1.076	2.931	1.518 - 5.660	0.0014

Table 4: CLMM with Confidence showing significant (p<0.05) effects from Sentiment Accuracy.

lower than just adding the second NMT.

For entity accuracy (Table 2), we see a significant (p<0.05) improvement from adding RBMT ($exp(\beta)$=1.74, CI=0.186-0.922, p=0.0089). Based on this $exp(\beta)$, an analyst with 3:1 odds of being either iffy or right would increase their odds to about 5.2:1. We see a similar effect size for adding a second NMT, but it is not statistically significant, and the 95% confidence interval ranges from a detrimental 0.7 to a dramatic odds improvement of 3.6, so we cannot draw conclusions on the effect of a second NMT on entity accuracy. Once again, we see a negative β for the interaction between adding a second NMT and RBMT, although it is not significant.

For sentiment accuracy (Table 3), we see no statistically significant effects with adding a second NMT or RBMT (p>0.1). Sentiment is the deepest level of comprehension in this user study, so it was the least likely to be improved with the addition of a second NMT and/or RBMT. Sentiment judgment is beyond the scope of typical MT-enabled triage tasks, and these results show that adding a second NMT and/or RBMT does not improve accuracy reliably enough to suggest that the scope of MT-enabled triage should be expanded to include tasks at the level of sentiment judgment without oversight by analysts that know the language.

4.4 Effects of Interventions on Confidence

In addition to measuring accuracy, we also track self-declared user confidence. To assess the impact of adding RBMT and/or a second NMT on analyst confidence, we fit four additional models.

Coefficient	β	Odds Ratio	Confidence Interval	p
(Intercept)	1.665	5.287	2.930 - 9.545	3.2e-8
2-NMT	0.577	1.781	1.053 - 3.013	0.0315
w/ RBMT	0.445	1.560	1.038 - 2.345	0.0322
2-NMT+RBMT	-0.685	0.504	0.284 - 0.895	0.0194
Confidence	1.176	3.241	1.898 - 5.534	1.7e-5

Table 5: GLMM for Relevance Accuracy with Confidence, indicating well-calibrated Confidence.

Coefficient	β	$exp(\beta)$	Confidence Interval	p
2-NMT	0.283	1.327	0.602 - 2.924	0.4823
w/ RBMT	0.549	1.732	1.194 - 2.512	0.0038
2-NMT+RBMT	1.181	3.258	1.194 - 5.409	4.9e-6
Confidence	-0.514	0.597	0.357 - 1.001	0.0503

Table 6: CLMM for Entity Accuracy with Confidence, showing significant effects (p<0.05) from RBMT and interaction with NMT and RBMT.

Coefficient	β	$exp(\beta)$	Confidence Interval	p
2-NMT	0.246	1.278	0.707 - 2.311	0.4162
w/ RBMT	0.233	1.263	0.913 - 1.746	0.1584
2-NMT+RBMT	-0.353	0.702	0.448 - 1.101	0.1233
Confidence	1.321	3.748	2.420 - 5.800	3.2e-9

Table 7: CLMM for Sentiment Accuracy with Confidence, indicating well-calibrated confidence.

Following the pattern of the previous models, we fit a cumulative link mixed effects model (CLMM) with confidence as the response variable and second NMT, RBMT, and their interaction as fixed variables. We also added relevance accuracy, entity accuracy, and sentiment accuracy as fixed variables. This model shows whether each of these features (presence of each intervention and each type of accuracy) is a good predictor of the user's confidence.

As shown in Table 4, we observe a large increase in odds of higher user confidence from adding a second NMT output ($exp(\beta)$ =2.86, CI=1.098 - 7.466, p=0.032).Adding RBMT does not have a significant effect, and no significant interaction is observed. Relevance and Entity accuracy do not have a significant effect on user confidence, but Sentiment accuracy has a large statistically significant effect ($exp(\beta)$=2.93, CI=1.518 - 5.660, p=0.008), nearly tripling the odds of higher confidence with higher sentiment accuracy. This may indicate that sentiment judgment was front-of-mind when users chose their confidence level.

If analyst confidence is well-calibrated with analyst accuracy, it should be true that not only is accuracy a strong predictor of confidence but confidence is also a strong predictor of accuracy. Given that sentiment accuracy is a stronger predictor of analyst confidence than the presence of a second NMT and/or RBMT, we suspect that analyst confidence is reasonably well calibrated with at least sentiment accuracy. We can directly test this by adding confidence as another fixed effect in the relevance, entity, and sentiment accuracy models and comparing the results.

With confidence added to the relevance accuracy model as a fixed effect, we see minimal change in the effect of RBMT as shown in Table 5, but the odds ratio for adding a second NMT drops from 2.26 to only 1.78. Confidence is a strong predictor of relevance accuracy (OR=3.24, CI=1.898 - 5.534, p=4.95e-5), and the model with the confidence fixed effect is also a significantly better (p<0.01) model based on AIC (1310.8 vs 1335.8) and log-likelihood (-645.39 vs -661.92). The large confidence effect and model improvement suggest that analyst confidence is well-calibrated to relevance accuracy.

Adding confidence to the entity model (Table 6)

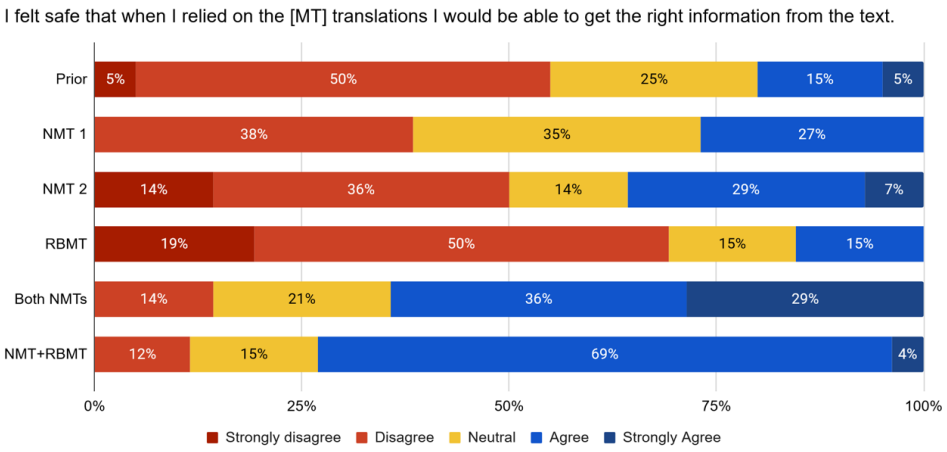


Figure 3: Participant responses to the safety item.

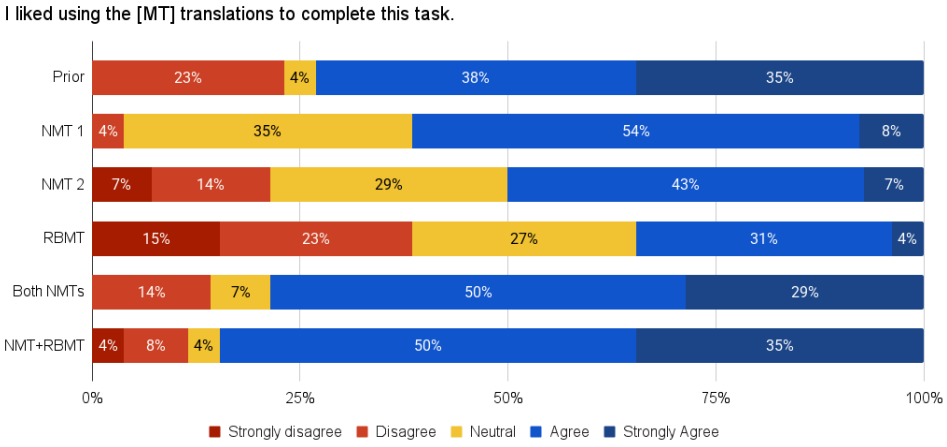


Figure 4: Participant responses to the likability item.

results in minimal change to the effects of adding a second NMT or RBMT, but confidence is as strong a predictor of entity accuracy as it was for Relevance accuracy (OR=3.26, CI=1.194 - 5.409, p=3.4e-5) and the entity accuracy model with confidence also demonstrates significant (p<0.01) improvements in AIC (1797.4 vs 1807.6) and log-likelihood (-882.68 vs -896.80) to those observed in the relevance model with the confidence fixed effect, suggesting that confidence is also well-calibrated with entity accuracy.

As with the original sentiment accuracy model, we see no significant effects from adding RBMT or a second NMT output (Table 7). Confidence is the only fixed effect to have a significant effect on sentiment accuracy (OR=3.75,CI=2.42-5.80, p=2.2e-8), verifying that just as sentiment accuracy is a strong predictor of confidence, confidence is also a strong predictor of sentiment accuracy. We also see significant (p<0.01) improvements to the AIC (2474.3 vs 2507.2) and log-likelihood (-1226.2 vs -1246.6) of the model from adding the confidence effect. This tells us that even when adding RBMT or a second NMT output does not affect accuracy, it also does not hurt confidence calibration.

4.5 User Feedback

For a mitigation to be effective, users must be willing to accept the resulting system. Key responses from the survey are the questions about likability and safety.

Figures 3 and 4 show how safe analysts felt when relying on the combinations of MT output and how much they liked using each combination. Less than 50% of participants agreed that they felt safe they would be able to get the right information from the text using the MT they typically have access to on-the-job or any one MT system from the user study.

With both NMTs, 65% of participants felt safe and 73% felt safe with NMT and RBMT. They liked having two NMT outputs (89%) but did not like using the NMT2 output as much as NMT1 (50% and 61%, respectively), and even though only 35% liked using RBMT, 85% liked using both NMT and RBMT. These seeming contradictions may be tied to how the analysts see themselves using the MT. Analysts may feel safe that they can get the right information because they believe they will be able to evaluate the information effectively. Similarly, analysts seem to like having access to RBMT as long as they have something to compare against. Prior work has indicated that IAs may be more likely than the general population to have an internal locus of control (Crouser et al., 2020), and that could explain their confidence that they will be able to take advantage of less-than-ideal MT output. Their open-ended responses give some insight as to how they use these combinations, with six analysts mentioning using Motrans for keyword spotting. As one analyst put it, "I used the literal translations very sparingly; mostly for the literal translation of a word, which I then plugged into the right spot of the neural translations."

5 Conclusions

We conducted a user study to establish a baseline level of IA performance on MT-enabled triage tasks and to measure the potential mitigating effects of a simple intervention, providing additional MT outputs. The user study found significant improvements in relevance judgment accuracy with output from two distinct NMT models and significant improvements in relevant entity identification with the addition of Motrans RBMT. The availability of additional MT outputs had little effect on analyst accuracy for the task that required the deepest comprehension of the text, identifying the sentiment towards the identified entity. Adding Motrans RBMT output had little effect on analyst confidence, but providing a second NMT output significantly improved it. This does not appear to be overconfidence, as confidence remained a strong predictor of accuracy across all three types of accuracy. Analysts also expressed a preference for seeing multiple MT outputs even when they felt that NMT1 provided better translations and praised the availability of multiple outputs in their open-ended post-task survey responses.

6 Recommendations and Future Work

Based on the analysts' preferences and the improvements in relevance judgment accuracy, we recommend that two MT outputs be displayed side-by-side wherever IAs conduct MT-enabled triage. RBMT such as Motrans, which can be rapidly updated with new named entities and technical terms, can help analysts with keyword spotting when the NMT misses them, but a second NMT may provide more benefit to relevance judgment overall. If it is practical to provide outputs from two NMT systems that are sufficiently different in model architecture and/or training data, users can benefit from the readability of the NMT while also gaining the ability to triangulate meaning between the two outputs. Some MT systems (including SYSTRAN) provide the ability to integrate terminology lists, which could replicate the entity recognition benefits of the RBMT system with the fluency of NMT.

However, we caution that despite the significant improvements to relevance judgment accuracy from providing multiple MT outputs, this should not be taken as evidence that these interventions will allow analysts to perform tasks using MT output that require higher levels of comprehension than triage. The lack of significant improvement in sentiment accuracy supports maintaining the status quo of not reporting off MT output without verification by a language-enabled analyst.

This study began before LLMs were available, leaving several open opportunities for future work. Rather than using two NMT models, a single LLM could be used to produce more than one translation, as Gero et al. (2024) did with a variety of sensemaking tasks. LLMs can also be prompted to post-edit (e.g., Xu et al., 2024; Raunak et al., 2023; Chen et al., 2024; Vidal et al., 2022; Ki and Carpuat, 2024) or provide quality estimation (e.g., Huang et al., 2024; Rei et al., 2023; Fernandes et al., 2023). Further work is needed to determine the optimal way to use these approaches to benefit MT-enabled triage use cases.

Additionally, more user testing is needed to determine ways to effectively display multiple translations of longer text. The benefits of the second MT output may be outweighed by the difficulty in actually comparing those outputs if long translations are just dumped into adjacent text boxes.

Acknowledgments

We gratefully acknowledge the invaluable quantitative analysis feedback and guidance from Susannah B. F. Paletz.

References

Tim Arango and Thomas Erdbrink. 2014. U.S. and Iran Both Attack ISIS, but Try Not to Look Like Allies. *The New York Times*.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 Shared Task on Quality Estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

Andrew S. Bowen, Clayton Thomas, and Carla E. Humud. 2022. Iran's Transfer of Weaponry to Russia for Use in Ukraine. CRS Report IN12042, Congressional Research Service.

Eleftheria Briakou, Navita Goyal, and Marine Carpuat. 2023. Explaining with Contrastive Phrasal Highlighting: A Case Study in Assisting Humans to Detect Translation Differences. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11220–11237, Singapore. Association for Computational Linguistics.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative Translation Refinement with Large Language Models. *arXiv preprint*. ArXiv:2306.03856 [cs].

R. Jordan Crouser, Alvitta Ottley, Kendra Swanson, and Ananda Montoly. 2020. Investigating the role of locus of control in moderating complex analytic workflows. *EuroVis 2020-Short Papers*.

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.

Ge Gao, Bin Xu, David C. Hau, Zheng Yao, Dan Cosley, and Susan R. Fussell. 2015. Two is Better Than One: Improving Multilingual Collaboration by Giving Two Machine Translation Outputs. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 852–863, Vancouver, BC, Canada. Association for Computing Machinery.

Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pages 1–21, New York, NY, USA. Association for Computing Machinery.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jiajun Chen, and Shujian Huang. 2024. Lost in the Source Language: How Large Language Models Evaluate the Quality of Machine Translation. *arXiv preprint*. ArXiv:2401.06568 [cs].

Carla E. Humud. 2023. Lebanese Hezbollah. CRS Report IF10703, Congressional Research Service.

Adam Tauman Kalai and Santosh S. Vempala. 2024. Calibrated Language Models Must Hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 160–171, Vancouver BC Canada. ACM.

Dayeon Ki and Marine Carpuat. 2024. Guiding Large Language Models to Post-Edit Machine Translation with Error Annotations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4253–4273, Mexico City, Mexico. 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, 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.

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation. In *Workshop on Neural Machine Translation*, Vancouver, BC. ArXiv: 1706.03872.

Yeskendir Koishekenov, Vassilina Nikoulina, and Alexandre Berard. 2022. Memory-efficient NLLB-200: Language-specific Expert Pruning of a Massively Multilingual Machine Translation Model. *arXiv preprint arXiv:2212.09811*.

Aviral Kumar and Sunita Sarawagi. 2019. Calibration of Encoder Decoder Models for Neural Machine Translation. *arXiv preprint*. ArXiv:1903.00802 [cs, stat].

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in Neural Machine Translation.

Daniel Licht, Cynthia Gao, Janice Lam, Francisco Guzman, Mona Diab, and Philipp Koehn. 2022. Consistent Human Evaluation of Machine Translation across Language Pairs. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 309–321, Orlando, USA. Association for Machine Translation in the Americas.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic Language Models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2022. Learning Confidence for Transformer-based Neural Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2353–2364, Dublin, Ireland. Association for Computational Linguistics.

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, Boston, MA. Association for Machine Translation in the Americas.

Marianna Martindale, Marine Carpuat, Kevin Duh, and Paul McNamee. 2019. Identifying Fluently Inadequate Output in Neural and Statistical Machine Translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 233–243, Dublin, Ireland. European Association for Machine Translation.

Marianna Martindale, Kevin Duh, and Marine Carpuat. 2021. Machine Translation Believability. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 88–95.

Marianna J. Martindale. 2012. Can Statistical Post-Editing with a Small Parallel Corpus Save a Weak MT Engine? In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2138–2142, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The Curious Case of Hallucinations in Neural Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

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.

Florence M. Reeder. 2000. At Your Service: Embedded MT As a Service. In *ANLP-NAACL 2000 Workshop: Embedded Machine Translation Systems*.

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.

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 689–730, Online. Association for Computational Linguistics.

Harini Suresh, Natalie Lao, and Ilaria Liccardi. 2020. Misplaced Trust: Measuring the Interference of Machine Learning in Human Decision-Making. In *Proceedings of the 12th ACM Conference on Web Science*, WebSci '20, pages 315–324, New York, NY, USA. Association for Computing Machinery.

SYSTRAN. 2021. Government Translation Solutions | SYSTRAN Technologies.

Blanca Vidal, Albert Llorens, and Juan Alonso. 2022. Automatic Post-Editing of MT Output Using Large Language Models. pages 84–106.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the Inference Calibration of Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3070–3079, Online. Association for Computational Linguistics.

Bin Xu, Ge Gao, Susan R. Fussell, and Dan Cosley. 2014. Improving machine translation by showing two outputs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3743–3746.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564.

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.

Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 295–305, New York, NY, USA. Association for Computing Machinery.

A Annotation Details

Screenshots of the annotation task are shown in Figures 5 and 6. Gold standard labels were assigned to items where both annotators agreed on the label. When annotators disagreed, the items were labeled ambiguous for the purpose of selecting items for the user study. Any ambiguous items that were eventually selected for the user study underwent a tie-breaking annotation where the original annotators were asked to come to an agreement on the final gold label. Interannotator agreement scores (Cohen's Kappa) before tie-breaking are shown in Table 9. Note that relevance applies to all items, but entity and sentiment apply only to items that both annotators labeled as relevant. Even before reconciliation, our annotators showed moderate to substantial agreement across the board and near-perfect agreement on entity and sentiment for Hezbollah and ISIS. The high level of agreement before reconciliation indicates that the annotators generally held the same understanding of the tasks and definitions, lending additional support to the reliability of the final reconciled labels.

The distribution of relevance, entity, and sentiment labels for comments in each task is shown in Table 8. In total, 32 out of the 61 comments included in the user study were labeled relevant. Because the comments were selected in threads, the relevant entities are not evenly distributed between tasks. All of the relevant comments in Russia/Ukraine Task A relate to Russia, compared to only half of the relevant comments in Russia/Ukraine Task B. On the reverse, only one comment in Russia/Ukraine Task A relates to Ukraine compared to all but one comment in Russia/Ukraine Task B. The Hezbollah comments are more evenly split, with three in Hezbollah/ISIS Task A and two in Hezbollah/ISIS task B, but the ISIS-related comments are almost all in Task B, with only one in Task A. The Hezbollah/ISIS tasks also contain fewer relevant comments overall compared to the Russia/Ukraine tasks. This difference is likely due to the recency of Russia's war in Ukraine at the time the comments were collected.

The annotation also included a human evaluation. For each comment displayed in the thread context, the language analysts rated the outputs of NLLB-200, SYSTRAN, and Motrans using a task-focused adaptation of the evaluation scales from Licht et al. (2022). Each quality level was given a descriptive label to emphasize that they are not meant to be equally distanced points in a range but rather descriptive quality levels. The descriptions of Licht et al. (2022)'s levels 4-5 were retained, but the descriptions of the first three labels were adapted to fit the levels of comprehension in the user study task. The lowest quality label was *MISS*, described as a translation that is so different from the meaning of the source text that a non-language-enabled

Figure 5: Screenshot of the relevance judgment and MT quality annotation view.

Figure 6: Screenshot of a completed comment annotation.

Task		Count	Relevant	Entity	Positive	Negative
Russia/ Ukraine	A	16	81.3% (13)	100% (13) 7.7% (1)	23.1% (3) 0.0% (0)	76.9% (10) 100.0% (1)
Russia/ Ukraine	B	14	57.1% (8)	50.0% (4) 87.5% (7)	0.0% (4) 57.1% (4)	100.0% (4) 42.9% (3)
Hezbollah/ ISIS	A	15	26.7% (4)	75.0% (3) 25.0% (1)	66.7% (2) 0.0% (0)	33.3% (1) 100.0% (1)
Hezbollah/ ISIS	B	16	43.8% (7)	28.6% (2) 71.4% (5)	100.0% (2) 0.0% (0)	0.0% (0) 100.0% (5)

Table 8: Distribution of gold standard relevance, entity, and sentiment labels for comments chosen for each task. *Relevant* indicates how often that entity was judged to be a relevant entity, and *Positive* and *Negative* indicate how often the sentiment towards that entity was positive or negative, respectively.

Label type	Russia/Ukraine	Hezbollah/ISIS	Combined
Relevance	0.537	0.566	0.574
Entity	0.684	0.834	0.740
Sentiment	0.679	0.972	0.799

Table 9: Annotator agreement (κ) on relevance, entity, and sentiment labels for our two annotators on the 556 comments (210 Russia/Ukraine; 346 Hezbollah/ISIS).

analyst would not be able to reliably make even a relevance judgment. The second level was *REL-ONLY*, described as translation quality sufficient to make a relevance judgment but with significant information missing or incorrect. The third level was *GIST*, described as translating critical information correctly but with less important information missing or incorrect. Levels 4 and 5 were labeled *GOOD* and *EXCELLENT* respectively. Translations below GIST quality (MISS or REL-ONLY) can be considered potentially misleading in this scenario.

Table 10 shows the percent of translations from each MT system that were given each of the labels and the percent that were potentially misleading (below GIST). Table 11 shows interannotator agreement (Kendall's Tau).

Motrans's lack of fluency is illustrated in the low percentage of translations that were at the GOOD or EXCELLENT level (9.53% and 5.94%, respectively), but its emphasis on adequacy is reflected in the smaller number of translations at the MISS level (10.79%) compared to NMT1 (17.45%) and NMT2 (13.13%). Because Motrans is rule-based MT, it cannot hallucinate or drop content as NMT models might, though it may mistranslate or leave words untranslated.

NMT2 (SYSTRAN) has the lowest percentage of Below GIST translations and the highest percentage of GOOD and EXCELLENT translations, suggesting that NMT2 might be a better match for these topics and this style than NMT1. However, these very specific domains (comments related to terrorist groups ISIS and Hezbollah and Russia's war in Ukraine) are only a small sample of domains that would need to be covered by a Persian-English MT system deployed to an intelligence analysis workforce. A multilingual model like NMT1 that demonstrates reasonable performance on a generic test set like FLORES may still be preferable as a baseline system, particularly if alternate NMT or RBMT proves beneficial in helping users overcome errors in the first NMT output.

MT	MISS	REL-ONLY	GIST	GOOD	EXCELLENT	Below GIST
NMT1	17.45%	31.12%	28.42%	13.13%	9.89%	48.56%
NMT2	13.13%	27.88%	32.19%	14.39%	12.41%	41.01%
Motrans	10.79%	38.67%	35.07%	9.53%	5.94%	49.46%

Table 10: Human quality judgments on all comment translations from NMT1 (NLLB-200), NMT2 (SYSTRAN), and Motrans.

Label type	Russia/Ukraine	Hezbollah/ISIS	Combined
NMT1	0.561	0.613	0.597
NMT2	0.666	0.561	0.602
RBMT	0.448	0.543	0.509
All	0.561	0.573	0.571

Table 11: Annotator agreement on MT quality labels (Kendall's tau) for our two annotators on the 556 comments (210 Russia/Ukraine; 346 Hezbollah/ISIS).

The GAMETRAPP project: Spanish scholars' perspectives and attitudes towards neural machine translation and post-editing

Cristina Toledo-Báez
Research Institute on Multilingual Language Technologies
University of Málaga
Spain
toledo@uma.es

Luis Carlos Marín-Navarro
Research Institute on Multilingual Language Technologies
University of Málaga
Spain
lmarin@uma.es

Abstract

The GAMETRAPP project (2022-2025), funded by the Spanish Ministry of Science and Innovation and led by the University of Málaga, aims to introduce and promote post-editing (PE) practices of machine-translated research abstracts among Spanish scholars. To this aim, the GAMETRAPP project is developing a gamified environment — specifically, an escape room—integrated into a responsive web app. As part of the design of both the gamified environment and the web app, this paper presents the results of a questionnaire distributed to Spanish scholars in order to explore their perspectives and attitudes towards neural machine translation (NMT) and PE. A total of 253 responses were collected from scholars affiliated with 42 Spanish public universities. A two-stage participant selection process was applied: the analysis focuses on scholars who self-reported a CEFR level of C1 or C2 in English proficiency. (n = 152), and, within this group, a comparison was conducted between scholars from linguistic disciplines (23%, n = 35) and those from non-linguistic disciplines (77%, n = 117). Statistically significant differences between these groups were identified using the Mann-Whitney U test in IBM SPSS. The results indicate a widespread and continued use of language technologies, particularly those related to NMT. However, only 34.2% of scholars from non-linguistic disciplines are familiar with PE as a concept, although 59.8% report that they do post-edit their scientific abstracts. Furthermore, 62.9% of scholars from linguistic disciplines and 47.9% from non-linguistic disciplines believe it is necessary to create an app that trains scholars in post-editing Spanish abstracts into English. Sentiment analysis conducted with Atlas.ti on the 29 qualitative responses to the open-ended question suggests overall neutral attitudes toward NMT and PE for both groups of scholars. In conclusion, while both groups engage with NMT tools, there is a clear need for training—especially among scholars from non-linguistic disciplines—to familiarize them with PE concepts and to help develop basic PE literacy skills.

1 Introduction and related work

Technology, particularly artificial intelligence (AI), plays a major role in shaping modern life, enabling numerous applications transforming various fields (Zhang et al., 2021). Translation technology has advanced significantly, driven by innovations like NMT (Sánchez Ramos and Rico Pérez, 2020) and pre-trained large language models (LLMs) (Brown et al., 2020). These AI-based methods have led to the development of a new generation of tools for translation and language services, including real-time language translation and communication through conversational chatbots such as ChatGPT (OpenAI, 2022; Jiang and Zhan, 2024; Rivas Ginel and Moorkens, 2024).

These myriads of resources and tools, combined with the growing globalization and interconnectivity, have led to NMT being deeply embedded in a wide range of professional, interpersonal, and social exchanges across the globe. As NMT is increasingly used by a wider number of people, initiatives such as the Machine Translation Literacy project (Bowker and Buitrago, 2019) and the MultiTrainNMT project (Kenny, 2022) have emerged with the aim of promoting NMT, training in NMT literacy, and raising awareness about the critical use that this technology requires.

Proceedings of Machine Translation Summit XX Volume 1, pages 608–618, June 23-27, 2025

One of the primary reasons for the growing demand for NMT arises from the increasing multilingualism in a society that requires seamless communication across multiple languages. However, this multilingualism clashes with the growing dominance of English as the lingua franca in research communication and international academic publishing (Curry and Lillis, 2019). The dominance of English, coupled with the global rise of the publish-or-perish culture in academia, is pushing scholars from both Anglophone and non-Anglophone countries to publish in English. The latter are currently referred to as English as an additional language (EAL) scholars (Zou et al., 2023) in the case of English for Research Publication Purposes (Flowerdew and Habibie, 2022).

The disparities resulting from the use of English as the dominant language in scholarly publishing are becoming more evident across various disciplines (Bowker, 2024). For instance, Amano et al. (2023) found that non-native English speakers spend considerably more time, effort, and money on reading and writing articles in English. To overcome the challenges of publishing in English and considering the improving quality of NMT output, scholars increasingly rely on MT—whether through NMT, LLMs, or chatbots—to write and translate their papers. Despite the high quality of results, it is still recognized that MT output generally requires PE to achieve a publishable quality. Defined, according to ISO 18587:2017, as "editing and correcting the output of a machine translation", the combination of NMT+PE in scholar communication has already been explored. For instance, Goulet et al. (2017) examined the use of NMT as a tool for composing academic texts in EAL, working with a group of ten researchers. Similarly, Parra Escartín et al. (2017) conducted a survey on the use of NMT by medical practitioners, subsequently analyzing their post-edits and assessing the final quality with the help of a professional proofreader. Other studies, such as those by O'Brien et al. (2018) and Parra Escartín and Goulet (2020), also conducted experiments aimed at exploring the relationship between NMT and PE, focusing on the quality and nature of the post-editing outcomes in each case.

Against the backdrop of scientific dissemination in English as EAL and the use of NMT+PE, the GAMETRAPP project (Toledo-Báez and Noriega-Santiáñez, 2024) is developing a web application that incorporates a gamified environment, specifically a virtual escape room, to introduce and promote the PE of research abstracts translated from Iberian Spanish to American English (L1 to EAL). While other applications, such as Kaninjo (Moorkens et al., 2016), have been developed to train users in PE, GAMETRAPP stands out by introducing gamification as an innovative strategy to engage users in the PE learning process. A key aspect when designing both a gamified environment and a web app is focusing on user needs and motivation (Herzig et al., 2015). Since the potential users of the GAMETRAPPP gamified environment and web app are Spanish scholars, a questionnaire was created and distributed to collect information on the methodology followed by scholars in Spain when writing and/or translating abstracts of their scientific publications.

For a participant-oriented study, it is common practice to use the term 'survey' to describe the study design, while the 'questionnaire' is seen as an instrument (Saldanha and O'Brien, 2014). A significant number of surveys and questionnaires about use of NMT and/or PE have already been conducted with professional translators (see Gaspari et al., 2015; Moorkens and O'Brien, 2017; Álvarez-Vidal et al., 2020; Canavese and Cadwell, 2024; Toledo-Báez, 2024, among others) and also with translation students (González Pastor, 2021; Zhang, 2023) and humanities students in general (Bowker, 2020; Dorst et al., 2022). However, aside from the aforementioned study by Parra Escartín et al. (2017), surveys and questionnaires regarding the use of NMT and/or PE by non- translators or non-linguists remain relatively limited. Anazawa et al. (2013) explored how Japanese nursing professionals used MT to access information from international journals. Their questionnaire results showed that more than half of participants found MT usable, and the study concluded that language proficiency is a key factor for the effective use of MT. Another study is Nurminen (2020), who interviewed nine Scandinavian patent professionals about their use of raw NMT in their professional practice, concluding that their use of NMT was both widespread and long-term.

The aim of this paper is to present the methodology and results of the questionnaire developed for the GAMETRAPP project, with a particular focus on the similarities and differences between scholars from linguistic and non-linguistic disciplines. It serves as a report on the user needs

analysis, reflecting the perspectives and attitudes of both groups of Spanish scholars toward NMT and PE.

2 Methodology

2.1 Research questions

Considering the introduction and the goal of creating a gamified environment and a web app to introduce and promote the PE of research abstracts among Spanish scholars, we present the following three research questions:

RQ1: How widespread is the use of NMT within Spanish scholars?

RQ2: How familiar are Spanish scholars with PE?

RQ3: To what extent is a training application for the PE of research abstracts from Spanish into English perceived as useful by Spanish scholars?

Both RQ1 and RQ2 will allow the as-is situation for Spanish scholars to be documented. RQ3 may provide relevant insights to the usefulness of an app for training on PE.

2.2 Questionnaire description

The questionnaire was designed using Google Forms and underwent a two-step validation process: first, by five experts—three scholars in Translation Studies and two scholars in Statistical Sciences— and, second, by the Ethics Committee for Experimentation at the University of Málaga. It was distributed in Spanish language[2] to scholars from all public and private universities in Spain. It was launched in mid-September 2024 and closed at the end of January 2025. To facilitate participation, various contact networks, LinkedIn, and mailing lists were used to invite Spanish scholars to complete the questionnaire.

A total of 253 responses were collected from scholars across 42 institutions, including Spanish public and private universities as well as research centers. Of these 42 institutions, 41 are public universities, representing approximately 98% of all public universities in Spain —demonstrating a strong level of representativeness. To analyze and present the questionnaire results, a two-stage participant selection process was applied. First, only participants who self-reported a Common European Framework of Reference for Languages (CEFR) level of C1 or C2 in English proficiency (n = 152) were selected, as these levels reflect advanced English language skills. Within this group, a further distinction was made between scholars from linguistic disciplines (n = 35, 23%) —specifically from the area of Linguistics, Translation, and Language Studies— and those from non-linguistic disciplines (n = 117, 77%), described all in Section 3.1. This distinction was made to explore the similarities and differences in the use of and familiarity with NMT and PE between scholars from linguistic and non-linguistic disciplines. Therefore, the analysis of this paper focuses on the responses of the 152 scholars who self-reported a CEFR level of C1 or C2 in English proficiency, comparing, in addition, responses from the 35 scholars from linguistic disciplines to those from the 117 scholars from non-linguistic disciplines.

The comparison between these two groups of scholars is further supported by a statistical significance test. Given that the results from the Kolmogorov–Smirnov test indicated a significant deviation from normality ($p < 0.001$), the non-parametric Mann-Whitney U test was employed in IBM SPSS to assess whether the differences between the two groups of scholars are statistically significant. A result is considered statistically significant if the p-value is less than 0.05 ($p < 0.05$).

The questionnaire consisted of two Sections. In the first one, all the demographic data of the participants were collected through 9 close-ended questions covering the following aspects:

a) general information about the participant (gender, age, position, years of experience, etc.)

b) areas of scientific production

c) mother tongue(s) and foreign/additional languages

d) self-reported English proficiency level

The second Section focuses on examining the methodology followed by Spanish scholars when writing and/or translating the abstracts of their scientific publications. This section includes a

[2] As the original questionnaire was drafted in Spanish, the English version is available at the following link: Access to the questionnaire.

significantly larger number of questions—18 in total—comprising 17 closed-ended and 1 open-ended item. The information collected covers the following aspects:

a) frequency of publication in English and Spanish

b) frequency of requests for an abstract in English

c) perceived ease of writing in and/or translating into English

d) use and perception of language technologies (NMT tools, online dictionaries, chatbots, parallel corpora, etc.)

e) use of external services of professional translators and/or post-editors

f) familiarity with PE concept

g) usefulness of an app to train on the PE of abstracts from Spanish into English

h) an open-ended item to gather voluntary additional comments on the questionnaire or any aspect of NMT or PE deemed relevant.

3 Results

3.1 Participants' background

The areas of scientific production for the 152 scholars selected (see Section 2.2.) are diverse, with some fields standing out more than others. Scholars from the linguistic disciplines—specifically within the area of Linguistics, Translation, and Language Studies —constitute the largest group (23%), followed by the scholars from Engineering and Architecture (21.9%), Social Sciences (12.6%), and Biomedical Sciences (7.7%). Other disciplines are represented to a lesser extent such as Law (6.6%), Sciences (6%), Maths and Physics (5.5%), Biology (5.5%), Chemistry (4.4%), Economics (4.4%), Natural Sciences (2.2%) and History, Geography and Arts (0.6%).

Concerning mother tongue(s), the predominant language is Spanish (85.6%), followed by other co-official languages of Spain, such as Catalan (8.2%) and Galician (1.2%). Other native languages reported include French (1.9%), Portuguese (1.9%), and English (1.2%). The most widely spoken foreign languages among respondents are English (67.8%), French (13.7%), and Italian (9.7%), followed by German (5.6%) and Portuguese (3.2%) at lower percentages.

3.2 Frequency of publication in English

As shown in Figure 1, scholars from linguistic disciplines are more frequently required to provide an abstract in English. A total of 54.3% (n = 19) report that they are 'Always' asked to provide an English abstract. The remaining respondents indicate that they are 'Usually' (31.4%, n = 11) or 'Sometimes' (11.4%, n = 4) asked to do so. The lowest percentage—2.9% (n = 1)— corresponds to those scholars who never publish in Spanish.

In contrast, responses from scholars in non-linguistic disciplines show a more balanced distribution. A total of 51.2% (n = 60) report being asked to provide an abstract in English, with equal proportions stating they are 'Usually' (25.6%, n = 30) or 'Always' (25.6%, n = 30) required to do so. Notably, 23.9% (n = 28) indicate that they do not publish in Spanish—a higher proportion than among scholars from linguistic disciplines—suggesting a greater need for translation or academic writing in English among non-linguists. The remaining respondents from non-linguistic disciplines report being less frequently asked for an English abstract: 19.7% (n = 23) are 'Sometimes' asked, and 5.1% (n = 6) are 'Never' asked to provide one. The comparison between two groups of scholars regarding the frequency of publication in English does not yield statistically significant results (p = 0.944).

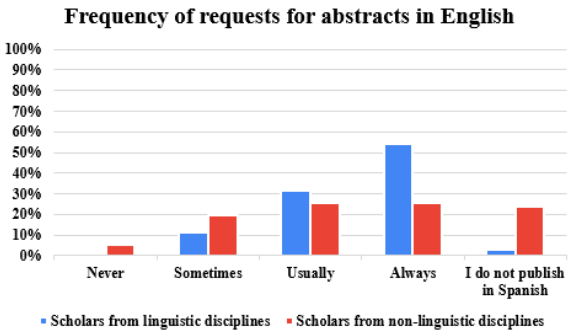


Figure 1: Frequency of requests for abstracts in English

When it comes to writing in English and/or translating abstracts into English, more than half of scholars from both linguistic and non-linguistic disciplines report that they find it not difficult. Specifically, 68.6% (n = 24) of scholars from linguistic disciplines and 53% (n = 62) of those from non-linguistic disciplines indicate no difficulty. A number of participants report only

minimal difficulties: 17.1% (n = 6) of linguistic scholars and 38.5% (n = 45) of non-linguistic scholars. The proportion of scholars who find it difficult is relatively small, with 14.3% (n = 5) from linguistic disciplines and 8.5% (n = 10) from non-linguistic disciplines reporting difficulty. No statistically significant differences were found between groups in relation to writing in English and/or translating abstracts into English (p = 0.241).

3.3 Perception and use of technological tools

Scholars from non-linguistic disciplines use NMT tools more frequently (78.6%, n = 92) than scholars from linguistic disciplines (71.4%, n = 25). Although the difference between the two groups is not statistically significant (p = 0.376), the results suggest that non-linguists tend to rely more heavily on NMT tools for translating their work. In contrast, linguists appear to be more critical of such tools and are more likely to use alternative methods.

Regarding the use of specific tools, scholars from linguistic disciplines show a stronger preference for online dictionaries, with a higher usage rate (60%, n = 21) compared to scholars from non-linguistic disciplines (50.4%, n = 59). Although the difference is not statistically significant (p = 0.321), this suggests that linguists tend to place greater emphasis on lexical precision and terminology accuracy.

Concerning the use of chatbots, a notable similarity is observed between the two groups: 48.6% (n = 17) of scholars from linguistic disciplines and 47% (n = 55) of scholars from non-linguistic disciplines. This balance, although it is not statistically significant (p = 0.871), suggests that the multi-disciplinary nature of these emerging conversational assistants—used not only for linguistic tasks but also for their interactive features—appeals equally to both linguistic and non-linguistic scholars. A notably higher proportion of linguists (31.4%, n = 11) use parallel corpora in contrast to non-linguists (19.7%, n = 23), highlighting that linguists are more inclined to work with corpora for comparative linguistic studies, ensuring terminological consistency, or validating translations, However, the differences between the two groups are still not statistically significant (p = 0.144). Other tools, such as Grammarly and IATE, are used exclusively by scholars from non-linguistic disciplines, with 3.4% (n = 4) using Grammarly and 0.9% (n = 1) using IATE. However, the differences observed in the data are not statistically significant (p = 0.269).

Only 6 scholars from linguistic disciplines (17.1%) and 8 from non-linguistic disciplines (6.8%) reported using no technological tools. The difference, while close to statistical significance (p = 0.065), is still not significant.

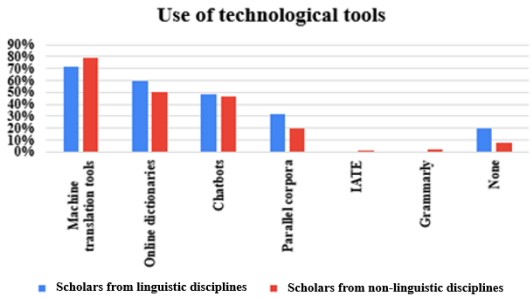


Figure 2: Use of technological tools

3.4 Knowledge and use of NMT and PE

According to the data obtained from the questionnaire analysis, it is evident that the concept of PE is largely unfamiliar to scholars outside of linguistic fields. Specifically, 65.8% (n = 77) of non-linguists are unaware of PE, compared to 34.2% (n = 40) who are familiar with the concept. In contrast, more than 90% (n = 33) of linguists are familiar with PE. This is the only variable with a statistically significant difference (p < 0.001), suggesting academic background plays a strong role in familiarity with PE. Linguists are significantly more likely to recognize or understand the concept than their non-linguistic counterparts.

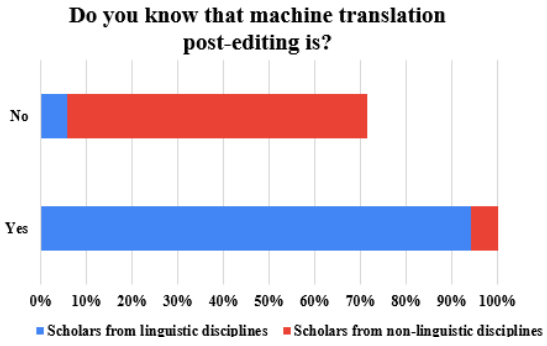


Figure 3: Do you know what machine translation post-editing is?

The data shown in Figure 4 below, based on experience with post-editing machine-generated translations, reveals that 88.6% (n = 31) of scholars from linguistic disciplines have post-edited a

machine-generated translation at some point, while 11.4% (n = 4) have not. In contrast, among scholars from non-linguistic disciplines, 73.5% (n = 86) have post-edited a machine-generated translation, while 26.5% (n = 31) have not been involved in this process. This is noteworthy, especially given that, as observed in Figure 3, more than 50% of these scholars are unfamiliar with the concept of post-editing. Although the statistical significance is close to 0.05 (p = 0.064), it remains nonexistent. Nonetheless, these results suggest that, while the majority of both groups have experience with post-editing, scholars from linguistic disciplines tend to have a higher rate of involvement in this activity, likely due to their deeper understanding of MT processes.

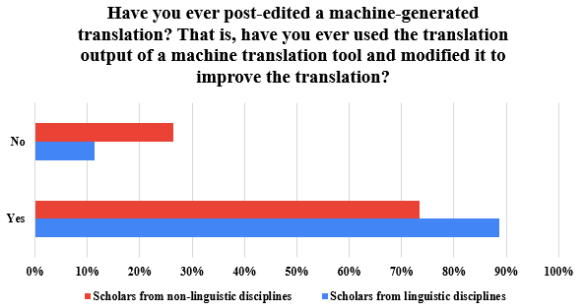


Figure 4: Have you ever used the machine-generated translation output of a machine translation tool and modified it in order to improve the result?

The statistics regarding PE of scientific abstracts reveal that, among scholars from linguistic disciplines, 74.3% (n = 26) have engaged in post-editing a scientific abstract, while 25.7% (n = 9) have not. In comparison, among scholars from linguistic disciplines, 59.8% (n = 70) have experience in post-editing scientific abstracts, while 40.2% (n = 47) have not participated in this activity. The statistical significance of this difference remains nonexistent (p = 0.121). These results suggest that, although both groups engage in post-editing scientific abstracts to a notable extent, scholars from linguistic disciplines have a higher rate of participation, indicating a potential correlation between linguistic knowledge and the practice of post-editing scientific texts.

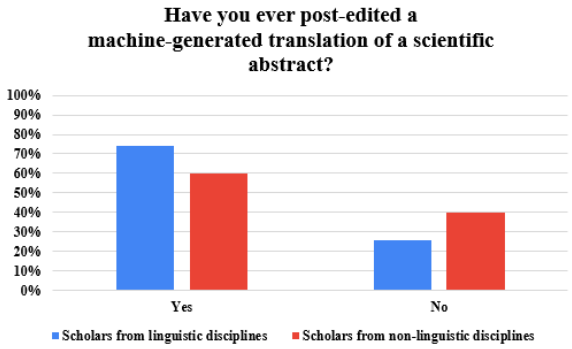


Figure 5: Have you ever post-edited a machine-generated translation of a scientific abstract?

3.5 Quality of NMT

The data reveal little differences in how the two groups of scholars rate the quality of NMT. Scholars from linguistic disciplines tend to rate the translation more leniently, with 57.1% (n = 20) deeming it 'Good', 21.4% (n = 6) rating it as 'Fair', and only 7.1% (n = 2) considering it 'Excellent'. Notably, there were no 'Poor' ratings from this group, suggesting they find the quality acceptable, though not outstanding. Scholars from non-linguistic disciplines also give a generally positive rating, with 70.8% (n = 51) considering it 'Good', and 6.9% (n = 5) rating it as 'Excellent', a percentage similar to that of linguistic scholars. Additionally, 22.2% (n = 16) rated it as 'Fair', and, like the linguistic group, no 'Poor' ratings were given. Overall, both groups rated NMT positively, and the slight differences between them were not statistically significant (p = 0.931).

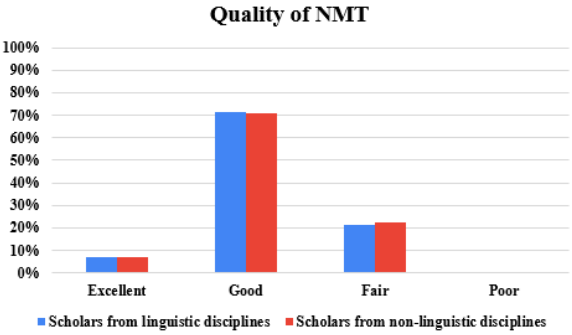


Figure 6: Quality of NMT

3.6 Usefulness of a training app for the PE of research abstracts from Spanish into English

Among scholars from linguistic disciplines, the majority (62.9%, n = 22) found an app designed to familiarize scholars with the PE of abstracts from Spanish into English to be 'Very useful', while a

smaller percentage (22.9%, n = 8) considered it 'Useful'. Only 14.3% (n = 5) rated it as 'Not useful', and no one marked it as 'Not useful at all'. In contrast, responses from scholars in non-linguistic disciplines were more varied: 47.9% (n = 56) found it 'Very useful', 30.8% (n = 36) rated it as 'Useful', 16.2% (n = 19) deemed it 'Not useful', and 5.1% (n = 6) considered it 'Not useful at all'. Although there is no statistical significance between the two groups (p = 0.112), the results suggest that linguists are more likely to view the potential app very positively, while scholars from non-linguistic disciplines rate it more neutrally, but still somewhat positively.

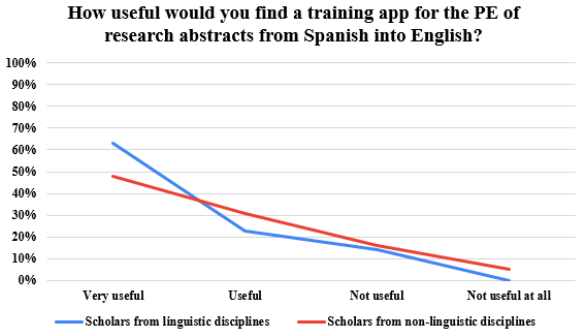


Figure 7: How useful would you find a training app for the PE of research abstracts from Spanish into English?

3.7 Sentiment analysis of open-ended question

A total of 29 qualitative responses (4 from scholars in linguistic disciplines and 25 from scholars in non-linguistic disciplines) were provided in response to the voluntary open-ended question, which sought additional comments on the questionnaire or any aspect of NMT or PE deemed relevant. A sentiment analysis was conducted with Atlas.ti in order to classify scholars' opinions. According to Luo et al. (2013), sentiment analysis, also known as opinion mining, uses natural language processing, computational linguistics and text analytics to identify and classify personal opinions in content sources, such as documents or sentences. The main goal of sentiment analysis is to determine the author's attitude on a specific topic or the general polarity of a document. The degree of sentiment will be determined by this polarity, i.e. a high positive score would indicate positive sentiment, while a low negative score would indicate negative sentiment. Neutral sentiment would be set at an intermediate score.

The sentiment analysis reveals that scholars from linguistic disciplines tend to comment more negatively: of the 4 comments in total, 3 express a negative attitude, and only 1 is neutral. Below we will present the English translation of a negative comment and a neutral comment chosen as examples:

Neutral comment: "1- I believe that chatbots (e.g., ChatGPT, Claude) that allow us to ask for help in translating and improving text, as opposed to traditional machine translation systems that lack the flexibility for interaction, already produce much better results. 2- In these cases, I wouldn't call it post-editing but rather focusing on how to write effectively (tone, precision, content, register, etc.). I think it's important to learn how to interact with Large Language Models (LLMs), including prompting, as well as developing the critical skill to read their responses and identify areas for improvement." (R210)[3].

Negative comment: "If they don't have any idea of L2 it won't do them any good. They need linguistic competence, and this post-editing gives you ideas on how to write, change a linker or something like that." (R14)[4].

In contrast, scholars from non-linguistic disciplines adopt a more neutral perspective: out of 25 comments, 14 are neutral, 9 are negative, and only 2 are positive. Below we will present the English translation of three comments (one positive, one negative and one neutral) chosen as examples:

[3] Original quote in Spanish: "1-Creo que hoy en día proporcionan ya resultados mucho mejores los chatbots (chatGPT, Claude, etc.) a los que podemos pedir que nos vayan ayudando a traducir y mejorar el texto, en lugar de utilizar sistemas de TA que no ofrecen la flexibilidad de 'interactuar'". 2-En estos casos, no hablaría tanto de poseditar, sino de cómo redactar bien (tono, precisión, contenido, registro, etc.). Creo que es importante aprender a interactuar con los LLM, el prompting, así como desarrollar la capacidad crítica para leer su respuesta e identificar los puntos que mejorar".

[4] Original quote in Spanish: "Si no tienen ni idea de la L2 no les va a servir de nada. Necesitan competencia lingüística y esta posedición al final lo que te da es ideas de redacción, cambiar algún linker o cosas del estilo."

Positive comment: "Pre-writing either in Spanish or English is essential to produce a good text in English." (140)[5].

Neutral comment: "I always consider it important that NMT should correct and help the researcher improve their English level or at least simplify the task of summarizing (with subsequent review). However, I do not believe that any technology should replace the need for a researcher to have a C1 level of English. Finally, it is important that a tool can be used proactively, rather than passively." (R7)[6].

Negative comment: "With AI tools I don't know if an application would be necessary." (R150)[7].

This difference among the two groups of scholars suggests that scholars from linguistic disciplines may be more critical of NMT and PE, likely due to their deeper familiarity with the challenges in these areas. Their views may reflect concerns about the limitations of NMT and the complexity of PE. On the other hand, scholars from non-linguistic disciplines seem to take a more relaxed approach, focusing less on the technical aspects of NMT and PE. Despite these differences, both groups share concerns about the effectiveness and quality of MT and PE. This contrast underscores the influence of educational and academic background on perceptions of technological developments in the field of translation.

4 Conclusions

This paper provides an overview of Spanish scholars' perspectives and attitudes towards NMT and PE through a questionnaire in which 253 Spanish scholars from 42 institutions participated. In order to analyze and present the questionnaire results, a two-stage participant selection process was applied. First, only participants who self-reported a CEFR level of C1 or C2 in English proficiency (n = 152) were selected, as these levels reflect advanced English language skills. Within this group, a further distinction was made between scholars from linguistic disciplines (23%, n = 35) and non-linguistic disciplines (77%, n = 117).

To address RQ1, data from our questionnaire indicate a widespread adoption of language technologies within the scientific community, with a particular preference for NMT tools among both scholars from non-linguistic disciplines (78.6%, n = 92) and scholars from linguistic disciplines (71.4%, n = 25). When compared to other studies, these results show a notable divergence. For instance, Moorkens and O'Brien (2017) found that only 18% of professional translators reported using NMT, and more than half of the respondents (56%) considered NMT to be "still a problematic technology". In contrast, the study by Canavese and Cadwell (2024) reported significantly higher usage rates, with 50.2% of respondents using NMT daily and 22.3% using it several times a week. The discrepancy may be partly explained by the six-year gap between them, reflecting the rapid evolution of NMT technologies. Nevertheless, neither study fully aligns with the findings of the present research, where 71.4% of scholars from linguistic disciplines reported using NMT.

When comparing our data on scholars from non-linguistic fields with those reported in Parra Escartín et al. (2017) —which focused on medical practitioners using NMT for academic writing support—we observe a strong similarity in NMT usage (68% in our study vs. 78.6% in theirs). The study by Anazawa et al. (2013), which also involved professionals in the health sciences, reports comparable findings: 65.8% of respondents use NMT to some extent, either 'Occasionally' (43.4%) or 'Always/almost always' (22.4%). These results align closely with those of Nurminen (2020), which highlight the widespread and long-term use of raw MT among respondents. Taken together, these findings suggest that reluctance and mistrust toward NMT are more pronounced among translation professionals and students than in other academic or professional fields.

Regarding RQ2, our study revealed a notable lack of awareness of PE, particularly among scholars from non-linguistic disciplines, with

[5] Original quote in Spanish: "La redacción previa ya sea en castellano o inglés es fundamental para tener un buen texto en inglés."

[6] Original quote in Spanish: "Siempre considero importante que la traducción automática debe corregir y ayudar al investigador a perfeccionar su nivel de inglés o en todo caso a simplificar la tarea de resumir (con revisión posterior). Pero no considero que ninguna tecnología deba suplir la necesidad de cualquier investigador de tener un C1 de inglés. En definitiva, es importante que se haga un uso proactivo de la herramienta y no tanto pasivo."

[7] Original quote in Spanish: "Con las herramientas de IA no sé si una aplicación sería necesaria."

65.8% reporting unfamiliarity with the concept. However, a majority of them (73.5%) indicated that they had engaged in PE at some point to improve NMT output. Furthermore, 59.8% of respondents from non-linguistic fields reported using PE specifically to enhance the quality of machine-translated scientific abstracts. The only prior study offering data on specific PE usage in scholarly communication is that of Parra Escartín and Goulet (2017: 260), which indicates that 26% of respondents use NMT "to obtain a preliminary English version they could subsequently post-edit." However, the study does not clarify how these scholars engage in PE, making direct comparison with our PE-related findings difficult and, in most cases, not feasible.

In relation to RQ3, the results suggest that scholars from linguistic disciplines are more likely to view a training app for the PE of abstracts from Spanish into English very positively. In contrast, scholars from non-linguistic disciplines tend to evaluate it more neutrally, though still with a generally positive outlook. As no previous studies have focused on the development of an app for PE, our results cannot be directly compared with existing research.

Our study has three main limitations that should be acknowledged. First, the total number of responses from scholars in linguistic disciplines was significantly smaller than that from scholars in non-linguistic fields. This imbalance was anticipated, as our focus was limited to a single area within the linguistic disciplines—namely, Linguistics, Translation, and Language Studies—compared to a total of 11 non-linguistic disciplines included in the study. Second, the overall response rate for the open-ended question was notably low, which can be attributed to its voluntary and unstructured nature. Third, the questionnaire did not offer respondents alternative methods for teaching basic PE skills. Only one option was presented, which limited the opportunity to compare it with other potential approaches to introducing and promoting PE.

The findings of this study point to at least two promising directions for future research. First, it would be valuable to explore alternative methods for teaching basic PE literacy skills, particularly to scholars from non-linguistic disciplines, as well as to other professionals or even the general public. GAMETRAPP introduces gamification as an innovative strategy to engage users in the PE learning process, and further studies will be conducted to assess its effectiveness. Second, it would be pertinent to investigate how PE literacy and skills evolve in the context of AI. The increasing use of NMT and LLMs for translation purposes could suggest that PE skills are becoming integrated into broader AI literacy (knowledge and skills) and AI competency (confidence and effectiveness) (Chiu et al., 2024). The integration of AI literacy and competency will become increasingly essential for effectively and responsibly navigating the digital transformation, necessitating particular emphasis on PE.

5 Acknowledgements

The GAMETRAPP project (TED2021-129789B-I00/AEI/10.13039/501100011033/Unión Europea NextGenerationEU/PRTR) is funded by the Spanish Ministry for Science and Innovation under the Ecological Transition and Digital Transition Call 2021.

6 References

Álvarez-Vidal, S., Oliver, A., and Badia, T. 2020. Post-editing for Professional Translators: Cheer or Fear? *Revista Tradumàtica. Tecnologies de la Traducciò,* 18: 49-69. https://doi.org/10.5565/rev/tradumatica.275

Amano, T., Ramírez-Castañeda, V., Berdejo-Espinola, V., Borokini, I., Chowdhury, S., Golivets, M., González-Trujillo, J. D. Montaño-Centellas, F., Paudel, K., White, R. L., and Veríssimo, D. 2023. The manifold costs of being a non-native English speaker in science. *PLoS Biology,* 21(7): e3002184. https://doi.org/10.1371/journal.pbio.3002184

Anazawa, R., Ishikawa, H., and Kiuchi, T. 2013. Use of online machine translation for nursing literature: A questionnaire-based survey. *Open Nursing Journal*, 7(1): 22–28. Available at: https://pubmed.ncbi.nlm.nih.gov/23459140/

Bowker, L. 2020. Chinese speakers' use of machine translation as an aid for scholarly writing in English: a review of the literature and a report on a pilot workshop on machine translation literacy. *Asia Pacific Translation and Intercultural Studies*, *7*(3): 288–298. https://doi.org/10.1080/23306343.2020.1805843

Bowker, L. 2024. Multilingualism in Scholarly Communication: How Far Can Technology Take Us and What Else Can We Do?". *The Journal of*

Electronic Publishing 27(1). https://doi.org/10.3998/jep.6262

Bowker, L., and Buitrago, J. 2019. *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Publishing.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyan, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radfor, A., Suutskever, I., and Amodei, D. 2020. Language Models are Few-Shot Learners. *Computer Science > Computation and Language*. https://doi.org/10.48550/arXiv.2005.14165

Canavese, P., and Cadwell, P. 2024. Translators' perspectives on machine translation uses and impacts in the Swiss Confederation: Navigating technological change in an institutional setting. In *Proceedings of the 25th Annual Conference of the European* Association for Machine Translation *(volume 2: Products and Projects)*, pages 347–359, Sheffield, United Kingdom. European Association for Machine Translation. https://aclanthology.org/2024.eamt-1.30.pdf (last accessed 02.07.2025).

Chiu, T. K. F., Ahmad, Z., Ismailov, M., and Sanusi, I. T. 2024. What are artificial intelligence literacy and competency? A comprehensive framework to support them. *Computers and Education Open,* 6. https://doi.org/10.1016/j.caeo.2024.100171.

Curry, M. J., and Lillis, T. 2019. Unpacking the lore on multilingual scholars publishing in English: A discussion paper. *Publications*, 7(2): 1-14. https://doi.org/10.3390/publications7020027

Dorst, A. G., Valdez, S., and Bouman, H. 2022. Machine translation in the multilingual classroom. How, when and why do humanities students at a Dutch university use machine translation? *Translation and Translanguaging in Multilingual Contexts*, 8(1): 49-66. https://doi.org/10.1075/ttmc.00080.dor

Flowerdew, J., and Habibie, P. 2022. *Introducing English for Research Publication Purposes*. Routledge.

Gaspari, F., Almaghout, H., and Doherty, S. 2015. A survey of machine translation competences: Insights for translation technology educators and practitioners. *Perspectives: Studies in Translatology*, 23(3): 333-358. https://doi.org/10.1080/0907676X.2014.979842

González Pastor, D. 2021. Introducing Machine Translation in the Translation Classroom: A Survey on Students' Attitudes and Perceptions. *Revista Tradumàtica. Tecnologies de la Traducciò*, 19: 47-65. https://doi.org/10.5565/rev/tradumatica.273

Goulet, M. J., Simard, M., Parra Escartín, C., and O'Brien, S. 2017. Using machine translation for academic writing in English as a second language: Results of an exploratory study on linguistic quality. *A. sp. Anglais de spécialité,* 72: 5-28. https://doi.org/10.4000/asp.5045

Herzig, P., Ameling, M., Wolf, B., and Schill, A. 2015. Implement gamification: requirements and gamification platforms. In *Gamification in education and business*, pages 431-450. Springer.

Jian, Z., and Zhang, Z. 2024. Can ChatGPT Rival Neural Machine Translation? A Comparative Study. Available at: https://arxiv.org/html/2401.05176v1

Kenny, D. 2022. *Machine translation for everyone: Empowering users in the age of artificial intelligence*. (Translation and Multilingual Natural Processing 18). Language Science Press.

Luo, T., Che, S., Xu, G., and Zhou, J. 2013. *Trust-based Collective View Prediction*. Springer.

Moorkens, J., and O'Brien, S. 2017. Assessing user interface needs of post-editors of machine translation. In *Human issues in translation technology*, pages 127-148. Routledge.

Moorkens, J., O'Brien, S., and Vreeke, J. 2016. Developing and testing Kanjingo: A mobile app for post-editing. *Revista Tradumàtica,* 14: 58-66. https://doi.org/10.5565/rev/tradumatica.168

Nurminen, M. 2020. Raw Machine Translation Use by Patent Professionals. A case of distributed cognition, *Translation, Cognition & Behavior*, 3(1): 100–121 https://doi.org/10.1075/tcb.00036.nur

O'Brien, S., Simard M., and Goulet M. 2018. Machine Translation and Self-Post-Editing for Academic Writing Support: Quality Explorations. In *Translation Quality Assessment. Machine Translation: Technologies and Applications,* pages 237–262. Springer.

OpenAI. 2022. *Introducing ChatGPT*. Website. Available at: https://openai.com/index/chatgpt/

Parra Escartín, C., and Goulet M. J. 2020. When the PostEditor is not a Translator: Can machine translation be post-edited by academics to prepare their publications in English? In *Translation Revision and Post-Editing,* pages 89–106. Routledge.

Parra Escartín, C., O'Brien, S., Goulet, M. J., and Simard, M. 2017. Machine Translation as an Academic Writing Aid for Medical Practitioners. In *Proceedings of MT Summit XVI*, 254–267, Nagoya, Japan.

Rivas Ginel, M. I., and Moorkens, J. 2024. A year of ChatGPT: translators' attitudes and degree of adoption. *Revista Tradumàtica*, 22: 258-275. https://doi.org/10.5565/rev/tradumatica.369

Saldanha, G., and O'Brien, S. 2014. *Research methodologies in translation studies*. Routledge.

Sánchez Ramos, M. M., and Rico Pérez, C. 2020. *Traducción automática. Conceptos clave, procesos de evaluación y técnicas de posedición*. Comares, Granada.

Toledo-Báez, C. 2024. Posedición y paridad humano-máquina en traducción automática neuronal: Un estudio empírico desde la traducción profesional. *Lebende Sprachen,* 69(2): 434-463. https://doi.org/10.1515/les-2024-0003

Toledo-Báez, C., and Noriega-Santiáñez, L. 2024. GAMETRAPP project in progress: Designing a gamified environment for post-editing research abstracts. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation*, pages 18–20, Sheffield, UK. European Association for Machine Translation. https://aclanthology.org/2024.eamt-2.10.pdf

Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J. 2021. Dive into Deep Learning. arXiv:2106.11342v5. https://doi.org/10.48550/arXiv.2106.11342

Zhang, J. 2023. Exploring undergraduate translation students' perceptions towards machine translation: A qualitative questionnaire survey. *Proceedings of Machine Translation Summit XIX*, Vol. 2: Users Track, pages 1–10. September 4–8, 2023, Macau SAR, China. https://aclanthology.org/2023.mtsummit-users.1.pdf (last accessed 02.06.2025).

Zou, C., Gong, W., and Li, P. 2023. Using online machine translation in international scholarly writing and publishing: A longitudinal case of a Chinese engineering scholar. *Learned Publishing*, 36(4): 585-595. https://doi.org/10.1002/leap.1565

Using Translation Techniques to Characterize MT Outputs

Sergi Alvarez-Vidal, Maria de Campo, Christian Olalla-Soler, Pilar Sánchez-Gijón
Universitat Autònoma de Barcelona (UAB)
{sergi.alvarez, maria.docampo, christian.olalla, pilar.sanchez.gijon}@uab.cat

Abstract

While current neural machine translation (NMT) and generative pre-trained transformer (GPT) models improve fluency and context awareness, they struggle with creative texts, where figurative language and stylistic choices are crucial. Current evaluation methods fail to capture these nuances, which require a more descriptive approach. We propose a taxonomy based on translation techniques to assess machine-generated translations more comprehensively. The pilot study we conducted comparing human and machine-produced translations reveals that human translations employ a wider range of techniques, enhancing naturalness and cultural adaptation. NMT and GPT models, even with prompting, tend to simplify content and introduce accuracy errors. Our findings highlight the need for refined frameworks that consider stylistic and contextual accuracy, ultimately bridging the gap between human and machine translation performance.

1 Introduction

Rapid advancements in neural machine translation (NMT) and generative pre-trained transformer (GPT) models have significantly improved the quality of machine-generated translations in recent years. In many cases, these models achieve an output product that closely resembles human translation (Jiao et al., 2023; Wang et al., 2023), making it increasingly difficult to describe or evaluate their performance using traditional metrics. Although early claims suggested that machine translation (MT) had reached parity with human translation (Hassan et al., 2018), subsequent studies have challenged these assertions, underscoring the persistent difficulties of evaluating machine-generated output in a way that captures their full complexity (Toral et al., 2018; Läubli et al., 2020). These debates further emphasize the limitations of current assessment methods, particularly their inability to account for the contextual and stylistic nuances (Wang et al., 2024) that professional translators consider essential.

One of the most pressing challenges in this context is the translation of creative texts, such as literature and marketing content. Unlike technical or informational texts, which often follow predictable structures and terminology, creative texts rely heavily on figurative language, including irony, metaphor, and ambiguous phrasing. These elements often lead to overly literal, word-for-word translations that do not convey the intended meaning in the machine-translated text (Guerberof-Arenas and Toral, 2020). Although current GPT models offer notable improvements by considering broader contextual relationships in sentences (Castilho et al., 2023), they still struggle with the complexities of creative expression.

Current evaluations are usually based on automatic metrics such as BLEU (Papineni et al., 2002) or COMET (Rei et al., 2020), or on manual evaluations that produce a list of errors and their severity, such as the MQM taxonomy (Lommel et al., 2014). However, these metrics mainly focus on the traditional accuracy and fluency paradigms, which do not account for any stylistic variation. Recent research has even shown that the inclusion of machine-translated texts in test data can significantly affect the results of evaluation outcomes. For example, Graham et al. (2020) found that MT systems may appear to perform better or worse depending on the nature of the test data.

Thus, we need to explore alternative approaches to describe and assess these texts in accordance with the contexts in which they are intended to be used. In contexts where both the conveyed information and the expressive or persuasive function of the text are essential, human translators frequently

Proceedings of Machine Translation Summit XX Volume 1, pages 619–627, June 23-27, 2025

employ a range of techniques to help the target audience grasp the subtle nuances of the original text. These strategies ensure that not only the content, but also the intended impact of the text is effectively conveyed in the translated version. If texts translated using NMT and GPT models are employed in the same scenarios where human translators apply these techniques, it is worth considering whether these techniques can also serve to describe and, consequently, evaluate the quality of machine-generated translations.

In this context, Translation Studies provide a rich theoretical framework that can offer more nuanced descriptive criteria. Specifically, we develop a taxonomy partially based on the translation techniques defined by Molina and Hurtado (2002). Their framework categorizes the translation techniques employed by human translators, which can serve as a benchmark for describing machine translations at a deeper level.

Using translation techniques such as modulation, amplification and explicitation, our proposed method aims to capture the complexity of translation beyond literal equivalence, helping us describe machine translation outputs or *machine translationese*. This approach enables us to assess how well MT models handle pragmatic and linguistic challenges, including idiomatic expressions, register changes, and cultural adaptation, thus providing a more comprehensive understanding of their strengths and weaknesses.

The remainder of this paper is structured as follows. Section 2 reviews related work on *translationese* regarding MT, highlighting some of the key concepts of its characterization. Section 3 introduces the proposed framework based on translation techniques and its theoretical underpinnings. Section 4 presents the setup and methodology used to conduct a pilot study of this framework, followed by the results in Section 5. Finally, Section 6 concludes the paper and outlines future research directions.

2 Human Translation vs Machine Translation

The study of differences between translated texts and non-translated texts has long been a central focus of translation studies research, with early research identifying distinct linguistic features that describe what has been called *translationese*. Toury 2012 differentiates between the law of interference, which refers to the elements of the source text that are retained in the translation, and the law of growing standardization, which relates to the tendency to apply the norms of the target language and culture to the translation product. Thus, any final translation is the hybrid result of the application of both laws.

Chesterman (2004) makes a distinction between S-universals and T-universals. S-universals are features that can be traced back to the source text. T-universals, on the other hand, are features that should be studied by comparing translated texts to non-translated texts in the target language, using a comparable corpus. They include features such as simplification, untypical patterning, and underrepresentation of target-language-specific items.

Baker (1993) suggests there are several translation universals, which are linguistic features that tend to characterize translated texts regardless of the language pair or direction of translation. These include simplification, where translations exhibit reduced structural and lexical complexity; explicitation, the tendency to render implicit information more explicit; normalization, which aligns translations more closely with conventional target language norms; leveling-out, which results in reduced variation across different text types; and interference, where source language structures influence the target text.

Corpora have been used extensively to study *translationese*. For example, Corpas Pastor (2008) argues that translated texts include lower lexical diversity, shorter sentence structures, and increased explicitation. These tendencies emerge due to the translator's dual commitment to preserving source meaning while ensuring readability in the target language. Empirical studies using comparable corpora have consistently shown that *translationese* manifests across languages, regardless of the specific translation directions (Volansky et al., 2015).

Human translations and machine translations have also shown divergences at the morphosyntactic level. Luo et al. (2024) conduct a large-scale fine-grained comparative analysis across three language pairs and show MT is consistently more conservative than human translations, as it shows less morphosyntactic diversity, more convergent patterns, and more one-to-one alignments.

As MT technology advances, researchers have begun to investigate whether similar patterns can be detected in MT-generated texts and post-edited (PE) translations (Castilho and Resende, 2022;

Toral et al., 2018). Some studies suggest that PE texts inherit certain traits from raw MT output, such as reduced lexical diversity and terminological consistency that align more closely with machine-generated texts than with human translations. For instance, Vanmassenhove et al. (2021) identify a loss in lexical richness in MT output, which could subsequently influence the characteristics of post-edited texts. Toral (2019) finds that post-edited documents have lower lexical variety and lower lexical density than human translations. Moreover, sentence length and parts-of-speech in post-edited texts are more similar to the source language than those in human translations.

A study by Zhu et al. (2024) examines translation relations to identify differences between NMT and human translations. The findings reveal that NMT systems tend to rely more heavily on literal translations compared to human translators, especially in the use of semantic-level translation techniques. The advent of large language models (LLMs) and GPTs has introduced the concept of *generatese*, referring to the distinct linguistic patterns produced by these models during text generation tasks, including translation.

He et al. (2024) investigate whether LLMs can mimic human translation strategies by analyzing source sentences and inducing translation-related knowledge such as keywords and topics. Their research shows that while LLMs can exhibit human-like translation strategies, there are challenges to reducing errors such as hallucinations and mistranslations, which are often associated with *generatese*.

Comparative analyses between human translations and machine-generated texts have highlighted notable differences. A study by Chen et al. (2024) proposes an iterative prompting approach for LLMs to self-correct translations. Interestingly, while this method reduces string-based metric scores, neural metrics suggest comparable or improved quality. These refined translations achieve better fluency, although other challenges related to *generatese* still remain. Other studies also suggest that LLMs generate translations that deviate more from the source text than those produced by NMT models (Vilar et al., 2023; Raunak et al., 2023).

3 Framework of Translation Techniques

Translation techniques play a fundamental role in Translation Studies, serving as essential tools to analyze and understand the procedures by which translators achieve equivalence between source and target texts, and have long been studied by translation scholars (Vinay and Darbelnet, 1958; Newmark, 1981, 1988; Chuquet and Paillard, 1989; Molina and Albir, 2002; Gibová, 2012). These techniques provide a framework for systematically identifying and categorizing the choices translators make during the translation process to address linguistic, cultural, and contextual challenges. Their significance extends to various aspects of translation theory and practice, contributing to improving translation quality and the development of pedagogical approaches.

Translation techniques allow for a structured approach to evaluating translation choices by offering a set of predefined categories that describe how equivalence is achieved at the micro-textual level. This systematic analysis helps identify patterns in translator behavior, and to compare different translations of the same text. By distinguishing techniques, we can better understand how translators navigate linguistic and cultural differences.

However, there is no consensus in academia on the classification and nomenclature of translation techniques. Vinay and Darbelnet (1958) were the first to publish a classification of translation techniques with a clear methodological purpose. They defined seven basic procedures operating on three levels of style and classified them between literal and oblique.

Nida (1964) suggests three types of translation techniques: additions, subtractions and alterations. These techniques are used to adjust the form of the message to the characteristic structure of the target language, to produce semantically equivalent structures, to generate adequate stylistic equivalences, and to produce an equivalent communicative effect.

Newmark (1988) uses the term *procedures* to classify translation techniques proposed by comparative linguists. These include: recognized translation, where an already accepted term is used even if it is not the most precise; functional equivalence, which replaces a term with a culturally neutral expression plus a qualifier; and naturalization, which adapts a source language word to the phonetic and morphological norms of the target language. He also introduces translation labels for provisional translations, often literal in nature. Additionally, Newmark allows for combining multiple procedures (doubles, triples, etc.) and includes synonymy as a separate category.

Molina and Hurtado (2002) modify and expand

previous classifications. They isolate the concept of technique by focusing on the notion of functionality, situating it in relation to the text and the context. For our framework, we take into account previous research on MT-generated content (Sanchez-Gijón, 2024; Zhai et al., 2024) and we make an effort to group the different phenomena in order to simplify corpus annotation. We simplify the original set of 18 translation techniques and add *naturalness*, which should be understood as a habitual use of the language, free of grammatical errors, fluid in style, and without expressions that are strongly influenced by other languages (do Campo Bayón and Sánchez-Gijón, 2024). Below we define the translation techniques and illustrate them with some examples from the annotated segments of the pilot study detailed in Section 4, for the Catalan-English language pair:

- **Non-literal linguistic choices in the pursuit of naturalness** This technique involves a departure from the original text, showcasing creativity in form while maintaining the original content. The translator prioritizes fluency and idiomatic expression in the target language to achieve a natural-sounding result.

  CA: Ja devia tenir un senyal vermell a la cintura, però així que el vent m'havia sortit per la boca la cinta tornava a fer-me el martiri. [I must have already had a red mark on my waist, but as soon as the wind had left my mouth, the ribbon went back to tormenting me.]
  EN: I pictured the red weal round my waist, but the moment I started rushing and getting out of breath, the elastic sliced into me again.

- **Established equivalent** This refers to the use of pre-existing, widely accepted equivalents in the target language, such as titles of movies, books, or brand names. By opting for the established equivalent, the translator ensures coherence and consistency with conventional usage.

  CA: La meva reina, va dir [My queen, he said.]
  EN: He said, my darling.

- **Simplification** Simplification entails the reduction of information without omitting essential meaning. It includes generalization and linguistic compression, conveying the same message with fewer details. Example:

  CA: La cinta de goma a la cintura estrenyent, estrenyent (...) [The rubber band around my waist, tightening, tightening.]
  EN: The elastic cutting deep into my waist (...).

- **Omission** The omission technique involves deliberately leaving out specific information that may not be essential for the overall message. The resulting text remains functional and coherent despite the absence of the omitted element.

  CA: (...) i a cada banda de la cara la medalleta de l'orella. [and on each side of the face, the little medal on the ear.]
  EN: (...) and little medal-like ears.

- **Explicitation** This technique makes implicit details (whether linguistic or thematic) explicit in the target text. It can include clarifying pronouns based on the level of formality or providing additional gender markers. Example:

  CA: Tan petita i ja té promès? [so young and you already have a fiancé?]
  EN: 'Aren't you too young to have a fiancé?'

- **Amplification** Amplification involves adding or making explicit details that the original audience might infer naturally. This technique is particularly useful when cultural or contextual knowledge cannot be assumed in the target audience.

  CA: (...) i vinga riure [and he kept on laughing]
  EN: (...) and he laughed till he cried.

- **Adaptation** Adaptation consists of finding an equivalent expression in the target language and culture that serves a similar function, even if it is not an established term. This technique is central to the domestication strategy, making the text more accessible and relatable to the target audience.

  CA: (...) la meva mare morta i sense poder-me aconsellar [my mother dead and unable to advise me]
  EN: (...) my mother dead and gone and not around to give me advice

- **Fluency and accuracy errors** These errors occur when the translated text contains unnatural phrasing, awkward constructions, or

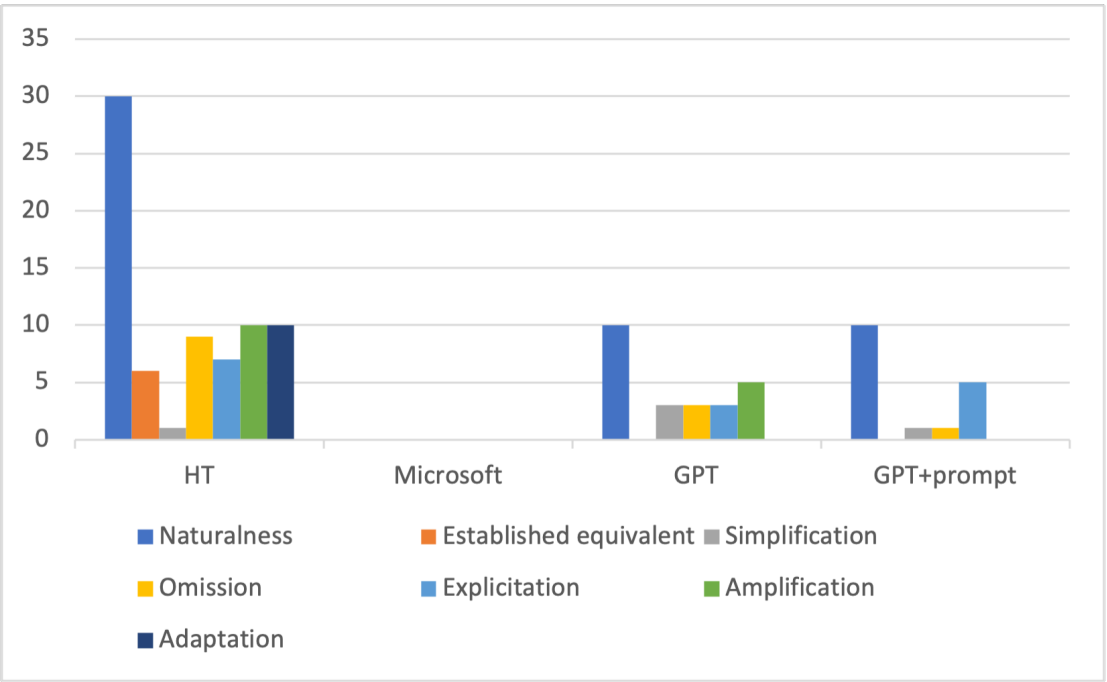


Figure 1: Use of techniques in the different translations

inaccuracies that may hinder comprehension or misrepresent the source text. They can include grammatical mistakes, stylistic inconsistencies, or mistranslations that affect the quality of the final output. These are the usual elements included in traditional evaluations and are incorporated in our annotation process to better understand the output translations in relation to usual evaluation techniques.

By applying these categories, we aim to gain deeper insights into the decision-making process of translators and the impact of various strategies on the final translated text.

4 Experimental setup and methodology

As an initial step following the selection of the translation techniques to be used for annotation, we decided to conduct a pilot study using one of the most renowned works in Catalan literature, *La Plaça del Diamant* by Mercè Rodoreda and its translation into English by Peter Bush in 2013. The novel was automatically segmented into sentences, and the first 60 segments were selected for the annotation process. We annotated the published translation into English and the translations produced by three MT engines. We used a NMT model (Microsoft Translator) and a GPT model (ChatGPT), as research shows these models translate broader contextual relationships across sentences better than NMT models (Castilho et al., 2023). Moreover, we used ChatGPT with a specific set of prompts to assess whether prompting techniques could improve the translation results for this type of text (Yamada, 2019; He, 2024).

We opted not to randomize the selection of segments, as the application of translation technique categories often relies on contextual references that extend beyond individual segments. Maintaining sequential order allowed us to preserve the coherence of the text and ensure that context-dependent techniques could be accurately identified and applied.

A relatively small number of segments was chosen for this pilot study, as its primary objective was twofold: first, to evaluate the relevance and applicability of the selected translation techniques; and second, to compare the results of the published human translation against raw machine translation (MT) outputs generated by NMT and GPT-based models with or without prompting techniques.

For each segment in the source language, four translations were annotated: (1) human translation, (2) Microsoft Translator translation, (3) ChatGPT translation without additional prompts, and (4) ChatGPT translation with specific prompts. For this version of ChatGPT, we introduced the following prompts in English, which described both the step-by-step actions followed by professional translators as well as some considerations regarding the

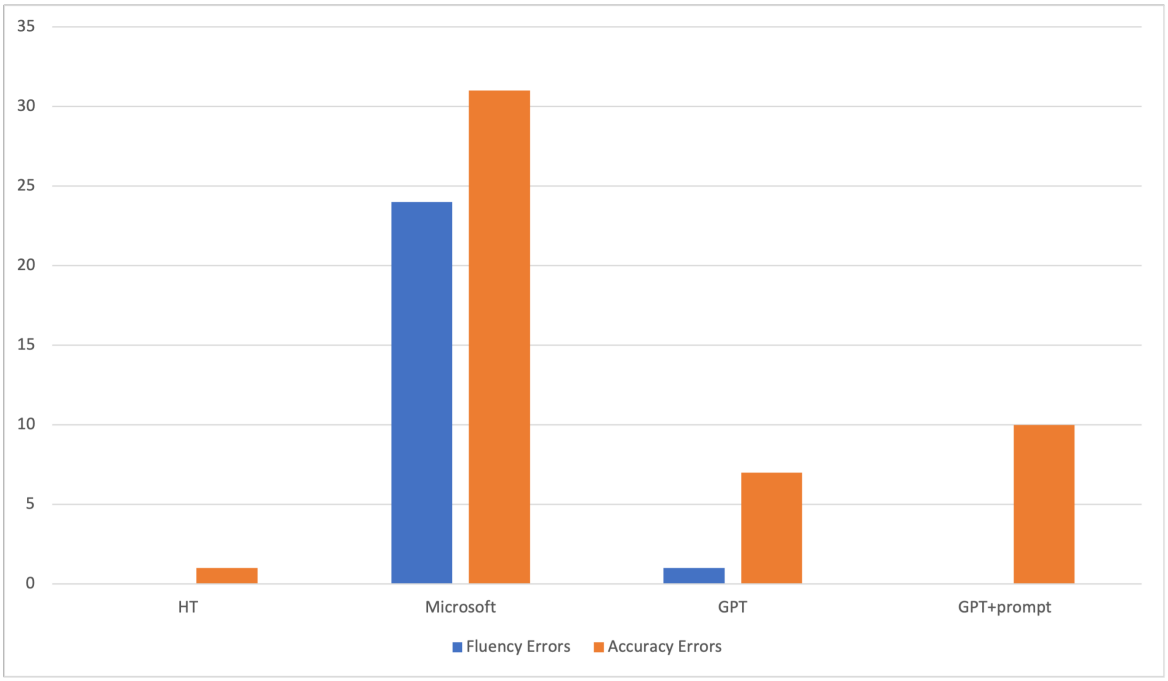


Figure 2: Accuracy and fluency errors

context and the text type:

Translate a literary text from Catalan (CA) to English (EN) while considering the cultural differences between the CA and EN readers. Follow a professional translator's strategy by considering the text's function, the cultural and social differences between the two audience groups, the author's style, and the text genre. Assume the EN reader is unfamiliar with CA culture, particularly regarding life in the 1950s in Barcelona and surrounding areas.

Steps to follow:

1. *Analyze the Original Text: Understand the text's purpose, the author's style, and specific cultural references unique to 1950s Barcelona.*
2. *Identify Cultural and Social Differences: Note key cultural elements that may need context or adaptation for an EN audience.*
3. *Translation Strategy: Adjust cultural references as needed to make them understandable without losing the text's authenticity. Maintain the original author's style and tone while ensuring it is accessible to an EN audience. Keep the genre conventions in mind to ensure the translated text aligns with expectations typical to that genre in English literature. Adapt for EN Readers: Provide additional context where necessary to enhance understanding of cultural nuances without altering the narrative.*
4. *Review and Revise: Ensure the final translation feels natural to an EN reader and accurately represents the original text's nuances.*

Output Format:

Provide the translated text in a natural and fluent English format, maintaining the original length as closely as possible while ensuring cultural clarity.

Each segment was annotated by two different annotators with previous experience in similar tasks. For the segments in which both annotators were not in agreement, a third annotator assessed the proposals and made a final decision.

5 Results

In Figure 1 we can see the results of the annotation process. Human translations include a higher number of translation techniques than any of the MT-produced translations, except for the simplification technique. In fact, this is one of the techniques that reduces source language information without any substitution or modification. However, human translation incorporates more omissions, which can be linked to the compensation process undertaken while translating, as in many other segments human translations incorporate techniques used to add more explicit information. It is also clear from the results that the NMT model (Microsoft) does

not use any of the translation techniques and thus produces more literal translations.

From the annotated techniques, we can highlight the use of naturalness in the human translation, which is the most frequently applied. In the search to produce a text that engages the target reader and has the same impact as the source reader, the translator makes decisions that move away from word-to-word translation and incorporate a creative component. Moreover, human translations also include increased use of the adaptation of the content (for example, with names of people and places) and amplification of certain elements to highlight them in the translation.

In Figure 2 we can see the results for accuracy and fluency for all output translations. All outputs contain a considerable high number of inaccuracies or translations which do not convey the meaning of the source text. Once again, the NMT model produces the highest number of fluency and accuracy errors, which are highly reduced in the case of ChatGPT. An interesting result is that the inclusion of prompts increases the number of accuracy errors. This could be linked to the effort made by ChatGPT to create more literary and creative content when the instructions explicitly indicate it. The creation of this type of translations seems to have as a side-effect the increased number of hallucinations or errors in the translations.

6 Conclusion and Future Work

The improved quality of the translations produced by the NMT and GPT models makes it increasingly difficult to distinguish them from human-produced texts. Current evaluation metrics fail to account for the stylistic and contextual nuances that are crucial in human translation. The challenge is particularly evident in the translation of creative texts, where figurative language plays a key role in meaning-making.

To address these limitations, we proposed a framework based on translation techniques, inspired by established models in Translation Studies. Our pilot study comparing human, NMT and GPT-produced translations of *La Plaça del Diamant* reveals significant differences in translation strategies. Human translators employ a wider variety of techniques, such as amplification, naturalness, and adaptation, that contribute to more natural, culturally appropriate, and stylistically coherent translations. In contrast, NMT and GPT models, even with targeted prompts, tend to simplify content, favoring more literal renderings that sometimes fail to capture the expressive function of the source text. While prompting techniques can make GPT translations appear more creative, they also introduce a higher number of accuracy errors, suggesting a higher introduction of hallucinations.

These findings reinforce the need for refined evaluation frameworks that move beyond traditional metrics to incorporate a deeper analysis of textual adaptation and stylistic effectiveness. By systematically categorizing translation strategies, our approach provides a more comprehensive way to assess how well machine translations handle complex linguistic and cultural challenges. Future research should build on this framework by expanding corpus size, and exploring automated annotation methods to improve scalability. Ultimately, integrating translation techniques into MT evaluation can offer a more human-centric perspective, bridging the gap between computational advancements and the nuanced decision-making process of professional translators.

References

Mona Baker. 1993. Corpus Linguistics and Translation Studies — Implications and Applications. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli, editors, *Text and Technology: In honour of John Sinclair*, page 233. John Benjamins Publishing Company.

Sheila Castilho, Clodagh Quinn Mallon, Rahel Meister, and Shengya Yue. 2023. Do online Machine Translation Systems Care for Context? What About a GPT Model? In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 393–417, Tampere, Finland. European Association for Machine Translation.

Sheila Castilho and Natália Resende. 2022. Post-Editese in Literary Translations. *Information*, 13(2).

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).

Andrew Chesterman. 2004. Beyond the particular. In Anna Mauranen and Pekka Kujamäki, editors, *Translation Universals: Do they exist?*, Benjamins Translation Library, pages 33–49. John Benjamins Publishing Company.

Hélène Chuquet and Michel Paillard. 1989. *Approche*

linguistique des problèmes de traduction anglais-français. Ophrys, Paris.

Gloria Corpas Pastor. 2008. *Investigar con corpus en traducción: los retos de un nuevo paradigma*, volume 49 of *Studien zur romanischen Sprachwissenschaft und interkulturellen Kommunikation*. Peter Lang.

María do Campo Bayón and Pilar Sánchez-Gijón. 2024. Evaluating NMT using the non-inferiority principle. *Natural Language Processing*, pages 1–20.

Katarína Gibová. 2012. Translation Procedures in the Non-literary and Literary Text Compared. *Bilingual Journal of Applied Linguistics*, 3(1):21–34.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical Power and Translationese in Machine Translation Evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81. Conference Name: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) Place: Online Publisher: Association for Computational Linguistics.

Ana Guerberof-Arenas and Antonio Toral. 2020. The impact of post-editing and machine translation on creativity and reading experience. *Translation Spaces*, 9(2):255–282. Publisher: John Benjamins Publishing Company.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv preprint*. ArXiv:1803.05567 [cs].

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).

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. *arXiv preprint*. ArXiv:2301.08745 [cs].

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Revista tradumàtica: traducció i tecnologies de la informació i la comunicació*, (12):455–463. Publisher: Departament de Traducció i d'Interpretació Section: Revista tradumàtica: traducció i tecnologies de la informació i la comunicació.

Jiaming Luo, Colin Cherry, and George Foster. 2024. To Diverge or Not to Diverge: A Morphosyntactic Perspective on Machine Translation vs Human Translation. *Transactions of the Association for Computational Linguistics*, 12:355–371.

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*, 67:653–672.

Lucía Molina and Amparo Hurtado Albir. 2002. Translation Techniques Revisited: A Dynamic and Functionalist Approach. *Meta: Journal des traducteurs*, 47(4):498–512.

Peter Newmark. 1981. *Approaches to Translation*. Pergamon Press, Oxford.

Peter Newmark. 1988. *A Textbook of Translation*. Prentice Hall, New York.

E.A. Nida. 1964. *Toward a Science of Translating with Special Reference to Principles and Procedures Involved in Bible Translating*. E.J.Brill, Leiden.

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.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs Produce Less Literal Translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. 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.

Sanchez-Gijón. 2024. Towards characterizing 'generatese. In *ATISA XI Conference*.

Antonio Toral. 2019. Post-editese: an Exacerbated Translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Antonio Toral, Martijn Wieling, and Andy Way. 2018. Post-editing Effort of a Novel With Statistical and Neural Machine Translation. *Frontiers in Digital Humanities*, 5.

Gideon Toury. 2012. *Descriptive Translation Studies – and beyond*. John Benjamins Publishing Company. Publication Title: btl.100.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. 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.

Jean-Paul Vinay and Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais: méthode de traduction*. Didier, Paris.

Vered Volansky, Noa Ordam, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1).

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.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Masaru Yamada. 2019. The impact of Google Neural Machine Translation on Post-editing by student translators. *The Journal of Specialised Translation*, pages 87–106.

Yuming Zhai, Gabriel Illouz, and Anne Vilnat. 2024. Annotation Guidelines of Translation Techniques for English-French. Technical report, LIMSI, CNRS.

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. _eprint: 2304.04675.

Author Index

The Machine Translation Summit XX organisers gratefully acknowledge the support from the following sponsors.

Platinum

Gold

Silver

Bronze

Supporters

Media

With the support of: