Team ACK at SemEval-2025 Task 2: Beyond Word-for-Word Machine Translation for English-Korean Pairs

Daniel Lee[*]
Adobe Inc.
dlee1@adobe.com

Harsh Sharma[*]
CU Boulder
harsh.sharma@colorado.edu

Jieun Han
KAIST
jieun_han@kaist.ac.kr

Sunny Jeong
New York University
sunny.jeong@nyu.edu

Alice Oh
KAIST
alice.oh@kaist.edu

Vered Shwartz
UBC
vshwartz@cs.ubc.ca

Abstract

Translating knowledge-intensive and entity-rich text between English and Korean requires transcreation to preserve language-specific and cultural nuances beyond literal, phonetic or word-for-word conversion. We evaluate 13 models (LLMs and MT models) using automatic metrics and human assessment by bilingual annotators. Our findings show LLMs outperform traditional MT systems but struggle with entity translation requiring cultural adaptation. By constructing an error taxonomy, we identify incorrect responses and entity name errors as key issues, with performance varying by entity type and popularity level. This work exposes gaps in automatic evaluation metrics and hope to enable future work in completing culturally-nuanced machine translation.

1 Introduction

Machine Translation (MT) has progressed significantly with the introduction of the transformer paradigm (Wang et al., 2023). Supervised machine translation models have improved their performance in challenging scenarios such as long-document translation and stylized translation (Lyu et al., 2024). Despite the success of these models in general translation, they still struggle to translate named entities which are culturally-naunced or language-specific (Xie et al., 2022), in other words, entities which are rooted in social, geographic, historical, and political contexts (Hershcovich et al., 2022). However, the emergence of self-supervised large language models (LLMs) and their zero-shot translation capabilities have shown to be a promising avenue to address these problems. Their ability to learn in-context enables new capabilities, such as terminology constrained translation unlike foundation approaches (Koshkin et al., 2024).

In this paper, we conduct a thorough analysis of English-to-Korean translation through the following:

- We conduct a comprehensive evaluation of 13 models (including LLMs and traditional MT models) on English-Korean translation pairs, focusing specifically on knowledge-intensive and entity-dense text.
- We complete a thorough human evaluation with bilingual annotators to construct a comprehensive error taxonomy.
- We reveal important gaps in automatic evaluation metrics (comparing BLEU, COMET, and M-ETA scores against human assessments), demonstrating that these metrics often fail to capture cultural and linguistic nuances in entity translation.

We hope this focused work on English-Korean motivates similar work in other domains, to further understand culturally-nuanced and language-specific translation.

2 Related Work

While MT has continued to improve from RNN-based models (Sutskever et al., 2014) to transformer-based models (Koishekenov et al., 2023; Tang et al., 2020; Zhu et al., 2024; Alves et al., 2023; Wang et al., 2023; Zaranis et al., 2024, *inter alia*), entity translation remains a significant challenge due to the need for both direct word-for-word conversion (*transliteration*) and contextual adaptation (*transcreation*) (Hershcovich et al., 2022). For example, the English query, "*What is the Rotten Tomatoes score of John Wick?*" should result in "*Rotten Tomatoes*" being translated to "로튼 토마토", the movies and TV review site, instead of "썩은 토마토", the literal translation meaning rotting fruit. While LLMs are a promising avenue to address these problems, they are primarily

[*] These authors contributed equally.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 2376–2388
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

trained on large-scale multilingual corpora with an English-centric bias. This results in them struggling to capture the nuanced sociocultural and historical contexts necessary for effective transcreation (Ponti et al., 2020). Efforts to enhance entity translation through retrieval-augmented generation (RAG) have been introduced, such as leveraging knowledge graphs (KGs) and structured databases. For example, KG-MT proposed using multilingual knowledge graphs to improve cultural adaptation in MT by providing contextually appropriate entity translations (Conia et al., 2024).

The complexities of English-Korean entity translation stem from fundamental linguistic and cultural differences between the two languages.

Transliteration is complicated by phonetic variations and word structure differences, while transcreation requires adapting names, idioms, and references to maintain cultural and linguistic naturalness (Pedersen, 2014; Díaz-Millón and Olvera-Lobo, 2023). Existing research has primarily focused on Western-centric language pairs, leaving English-Korean entity translation an area in need of further investigation (Kim and Choi, 2015; Kim et al., 2022). Given these challenges, improving LLM-driven MT requires context-sensitive modeling and culturally aware translation strategies. This study aims to bridge these gaps by evaluating state-of-the-art LLMs and multilingual MT models on entity-dense and knowledge-intensive texts, combining automatic evaluation metrics with human assessment to gain insights into translation quality.

3 Experimental Setup

To comprehensively evaluate the performance of LLMs on the task of MT, we consider 13 models including the most popular and best performing LLMs from OpenAI (GPT-4, GPT-4o, o1, o1-mini), Anthropic (Claude 3.5 Sonnet, 3.5 Haiku), Google (Gemini 1.5 Flash, 1.5 Pro), Meta (Llama3-8B), Grok (grok-2), and DeepSeek (R1-7B) and recent multilingual MT models (NLLB-200 and mBART-50) (OpenAI, 2024b,a; Anthropic, 2024; Gemini Team, 2024; xAI, 2024; DeepSeek-AI et al., 2025; Grattafiori et al., 2024; Koishekenov et al., 2023; Tang et al., 2020). With these models, we conduct an automatic evaluation with several metrics (Section 3.1) and an in-depth human evaluation (Section 3.2). For this evaluation, we use the task dataset provided by Conia et al. (2025) that was prepared from XC-Translate (Conia et al., 2024). XC-Translate is a multi-reference, human-curated dataset that is challenging due to its focus on translating cross-cultural texts containing entity names.

		Metrics		
Company	**Models**	*BLEU*	*COMET*	*M-ETA*
OpenAI	*o1*	0.3869	0.9196	0.3752
	o1 Mini	0.3830	0.9202	0.3306
	GPT-4o	0.3692	0.9087	0.3951
	GPT-4o Mini	0.3545	0.9046	0.2914
Anthropic	*Claude 3.5 Sonnet*	0.1961	0.8384	0.3969
	Claude 3.5 Haiku	0.1584	0.8056	0.2849
Google	*Gemini 1.5 Pro*	0.3810	0.9094	0.4833
	Gemini 1.5 Flash	0.2965	0.9081	0.3316
xAI	*Grok 2*	0.3808	0.9143	0.3514
DeepSeek	*DeepSeek R1*	0.0066	0.4895	0.0026
Meta	*Llama 3*	0.0327	0.5529	0.0563
	Mbart-50	0.1451	0.8702	0.0791
	NLLB-200	0.2195	0.8899	0.1663

Table 1: Automatic evaluation results on BLEU, COMET, and M-ETA for English-Korean translation. **Color key:** = Highest BLEU, = Highest COMET, = Highest M-ETA.

3.1 Automatic Evaluation

The automatic evaluation was conducted on the same 5,082 English-Korean pairs for each model, which comprised of several machine translation metrics including:

- **BLEU:** Measures the n-gram overlap between the translated text against the reference translations (Papineni et al., 2002).
- **COMET:** Predicts human judgments of machine translation quality using neural models (Rei et al., 2020).
- **M-ETA:** Measures the translation quality at the entity level (Conia et al., 2024).

We use a combination of the three automatic metrics augmented by human evaluation as they individually do not capture the nuances of machine translation. BLEU, while the industry standard and resource efficient, lacks strong correlation to human judgment (Callison-Burch et al., 2006). COMET better correlates with human judgments thanks to moving beyond n-grams to semantic understanding, yet it does not reveal fine-grained word-level insights (Kaffee et al., 2023) like culturally or language appropriate entities. M-ETA adds to both by focusing (only) on entity-level translation quality. Finally, human evaluation can detect translation errors not captured by the automatic metrics and provide in-depth feedback. Due

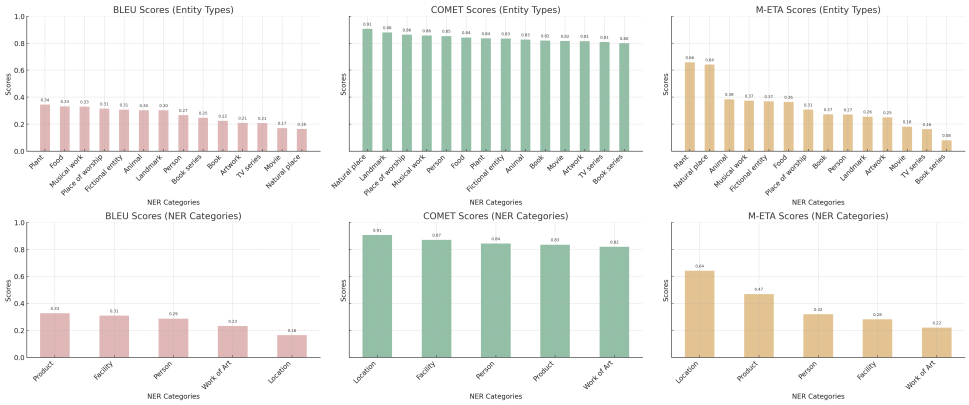

Figure 1: Average BLEU, COMET, M-ETA scores by entity types.

to its cost-intensive and time-consuming nature, we only manually evaluate a portion of the data. Used together, these metrics address the limitations proposed by each approach for a comprehensive analysis of machine translation quality.

3.2 Human Evaluation

For human evaluation, annotators were responsible for annotating 50 different English-Korean question pairs for each of the 13 models. They were tasked with evaluating (1) if the translation was correct, (2) which string in the English and Korean texts was mistranslated, and (3) a rationale for the mistranslation (Figure 4). To complete the task, we recruited two annotators with native fluency in English and Korean. Considering the importance of being able to understand and identify cultural and language-specific nuances in both English and Korean, annotators were required to have lived in South Korea and the USA for a minimum of 5 years each, and underwent a comprehensive interview to qualify for the task. Each annotator was compensated $150 for the completion of the entire task.

4 Results and Discussions

The automatic evaluation results presented in Table 1 summarize the BLEU, COMET, and M-ETA scores across all 13 models. For BLEU, o1 demonstrates superior performance, while o1-mini excels in COMET metrics and Gemini 1.5 Pro achieves the highest M-ETA score. These scores are closely followed by Grok-2 and GPT-4o across all metrics. Generally, most LLMs outperform traditional multilingual translation models such as MBART-50 and NLLB-200, with notable exceptions being DeepSeek R1, Llama 3, and both Claude 3.5 variants (Haiku and Sonnet).

To complement our automatic evaluation, we conducted human assessments to gain a more nuanced understanding of translation quality across models. Our analysis reveals that 459 out of 650 evaluated samples contain translation errors, with Grok 2 exhibiting the lowest error rate. Among these mistranslations, 266 cases involve incorrectly translated entities, which annotators identified by highlighting discrepancies between the English source and Korean target texts.

We further constructed a comprehensive error ontology adapted and expanded from (Popović, 2018) with annotator-provided explanations, illustrated in Table 2. The predominant error categories are "Incorrect Response" (308 pairs) and "Incorrect Entity Name" (266 pairs). "Incorrect Response", which encompasses behaviors unrelated to translation (e.g., answering questions rather than translating content), is most common, despite using identical prompts across all models as shown in Figure 2 due to its simple task. "Incorrect Entity Name" confirm that translating entities in knowledge-intensive and entity-dense texts remains particularly challenging. The primary failure modes involve literal, phonetic, or word-for-word translations that fail to capture the semantic content of the source text, demonstrating limited cross-lingual comprehension of entities. The definitions for each error label can be found in Table 5.

5 Further Analysis

We conduct a deeper, comprehensive analysis from our evaluation results to understand fine-grained insights in machine translation. In Section 5.1, we investigate whether entity popularity and typology influence translation quality. Section 5.2 explores

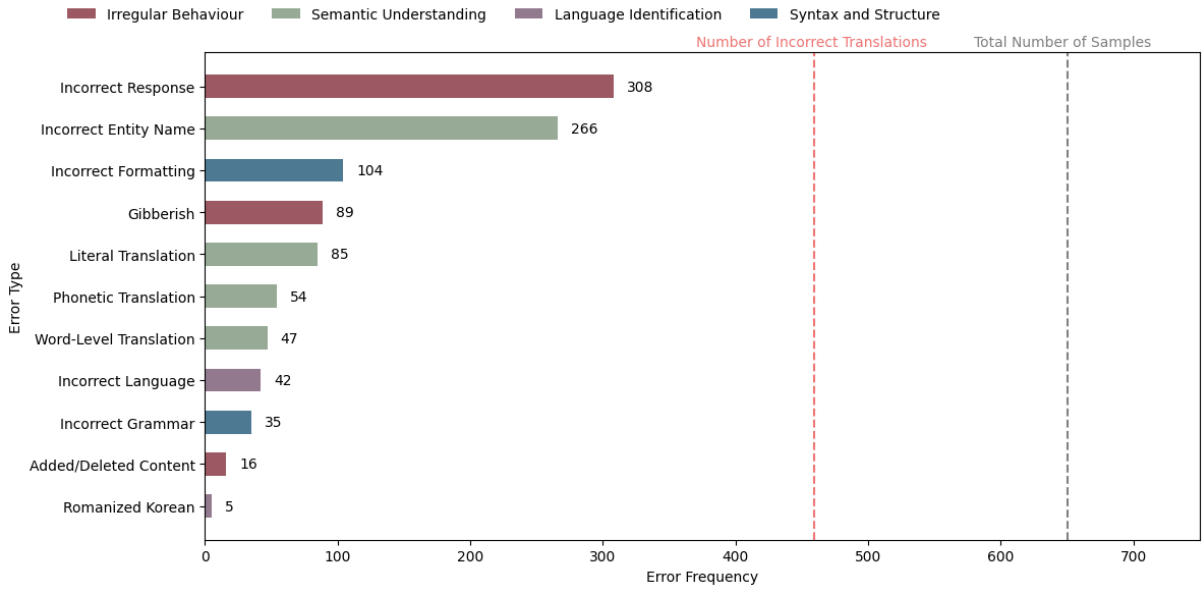

Figure 2: Frequency of errors per error taxonomy label.

the relationship between automatic metrics and human evaluations, providing insights into their complementary nature and potential discrepancies.

5.1 Impact of Entity Popularity and Type

We examine whether entity popularity influences machine translation quality, hypothesizing that frequently occurring entities in training data may yield better translations. Since the training corpora for these models are not publicly accessible, we use entity prominence as measured by Wikipedia page view statistics as a proxy for frequency in training data, operating under the assumption that widely-recognized entities are more likely to appear frequently in the data used to train these systems. To test this, we categorized entities from our dataset into five popularity segments based on their 2024 page view counts on Wikipedia: Low, Low-Mid, Mid, Mid-High, and High. As illustrated in Table 6, traditional metrics like BLEU and COMET remain relatively stable across these segments, with average scores of [0.26, 0.26, 0.25, 0.24, 0.27] and [0.84, 0.84, 0.83, 0.83, 0.83] respectively. However, M-ETA demonstrates a notable variation of 0.00224 across the popularity spectrum (Figure 3). Our findings suggest that while entity popularity impacts the translation quality of the entity itself, it does not significantly affect the translation of the surrounding sentence. Standard metrics such as BLEU and COMET fail to capture these nuances due to the relatively small token representation of entities within full sentences. This underscores the necessity for more fine-grained evaluation metrics that can assess culturally-nuanced and language-specific translation quality, rather than optimizing only for conventional translation metrics.

We further analyze performance by entity type (as categorized in Wikidata and standard named entity recognition (NER) types presented in (Tedeschi et al., 2021)) to investigate its influence on translation quality. Our results reveal performance disparities across different entity types in our dataset, with Plant and Natural place-related entities demonstrating higher performance across all 13 models. To distinguish between entity type effects and popularity level effects, we calculated the correlation between entity type performance and popularity scores 3. While the correlation patterns largely align with our previous findings, we observe several notable deviations.

Through qualitative analysis, we identify specific mechanisms by which entity types influence translation quality. For instance, entity type Book Series contains a higher concentration of names that could be simply literally, phonetically, or word-for-word translated, whereas entity type Plant and Natural place presents more challenges like requiring unique language-specific names. This suggests that translation difficulty is partially determined by the linguistic properties characteristic of specific entity categories, independent of their popularity or frequency in training data.

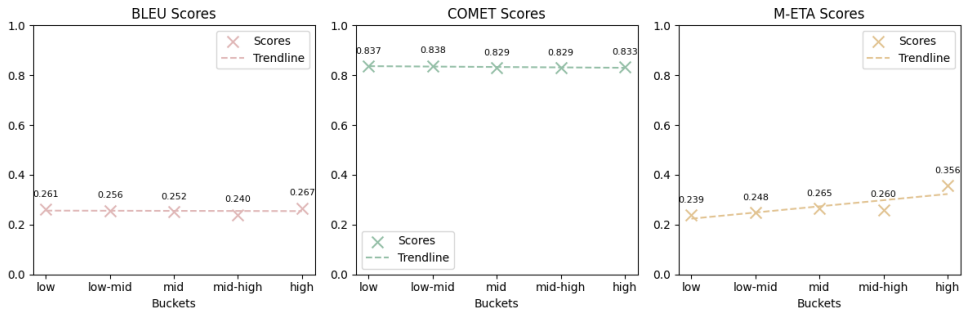

Figure 3: Average BLEU, COMET, M-ETA scores by popularity level.

5.2 Comparing Automatic and Human Evaluation

By comparing BLEU, COMET, and M-ETA scores with human evaluation results, we investigate the correlation between automatic metrics and human judgment to identify aspects of translation quality that may not be captured by computational assessment. Our analysis reveals that BLEU and COMET demonstrate a moderately positive correlation with binary human evaluations of translation correctness (i.e., whether the translated text preserves the meaning and comprehensibility of the source). Across 650 annotated samples, we observe a point-biserial correlation coefficient of 0.41 with a p-value of 3.54e-28, indicating a moderately strong alignment between automatic metrics and human assessment. M-ETA correctly identifies 88.7% samples containing entity translation errors according to human labels (Tate, 1954).

6 Conclusion and Future Work

In this paper, we evaluated 13 large language models and multilingual machine translation systems on their ability to handle culturally-nuanced and language-specific translation tasks across knowledge-intense and entity-dense questions from English to Korean. Our comprehensive assessment combined three automatic evaluation metrics with complementary human evaluations to thoroughly understand model performance. While our findings demonstrate that LLMs generally outperform traditional multilingual machine translation models, significant challenges remain, particularly regarding the appropriate transliteration versus transcreation of text. We hope this work encourages future research expanding beyond entity-dense and knowledge-intensive content to explore additional language pairs and text genres, ultimately informing targeted improvements in machine translation capabilities.

Limitations

In this section, we discuss some of the limitations of our work and how future research may be able to address them.

Language and Dialect Coverage. This paper focuses on a detailed analysis of Korean (Koreanic), for its morphologically complex, typologically different translation from English (Indo-European). However, it lacks an investigation into other language families like Romance (e.g., Spanish, French), Semitic (e.g., Arabic), Altaic (e.g., Turkish) and more. Future work should focus on related in-depth analysis on other languages and dialects, to develop a robust and generalizable understanding of errors in the culturally-nuanced and language-specific machine translation of text.

Error Ontology Coverage. We acknowledge the limitations of the proposed ontology, as our evaluation was restricted to a controlled set of question templates with a predefined entity pool from only English to Korean. A broader analysis of diverse text genres, such as long-form documents or narrative content, would likely reveal additional error categories. Furthermore, our current error classification system does not account for the varying degrees to which translation errors impact semantic comprehension across source and target languages, which means the framework inadequately captures important dimensions of translation quality.

Acknowledgements

We thank the organizers of SemEval 2025 Task 2: Entity-Aware Machine Translation (EA-MT) for creating a well-structured task. We also thank Simone Conia and Revanth Gangi Reddy for their valuable discussions and insightful feedback.

References

Duarte Alves, Nuno Guerreiro, Jo textasciitilde ao Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148. Association for Computational Linguistics.

Anthropic. 2024. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku. Accessed: 2025-04-27.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256. 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*, pages 16343–16360. Association for Computational Linguistics.

Simone Conia, Min Li, Roberto Navigli, and Saloni Potdar. 2025. SemEval-2025 task 2: Entity-aware machine translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Mar Díaz-Millón and María Dolores Olvera-Lobo. 2023. Towards a definition of transcreation: a systematic literature review. *Perspectives*, 31(2):347–364.

Google Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Technical report, Google DeepMind. Accessed: 2025-04-27.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang,

Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin

Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013. Association for Computational Linguistics.

Lucie-Aimée Kaffee, Russa Biswas, C. Maria Keet, Edlira Kalemi Vakaj, and Gerard de Melo. 2023. Multilingual Knowledge Graphs and Low-Resource Languages: A Review. *Transactions on Graph Data and Knowledge*, 1(1):10:1–10:19.

Gyeongmin Kim, Jinsung Kim, Junyoung Son, and Heuiseok Lim. 2022. KoCHET: A Korean cultural heritage corpus for entity-related tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3496–3505. International Committee on Computational Linguistics.

Youngsik Kim and Key-Sun Choi. 2015. Entity linking Korean text: An unsupervised learning approach using semantic relations. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 132–141. Association for Computational Linguistics.

Yeskendir Koishekenov, Alexandre Berard, and Vassilina Nikoulina. 2023. Memory-efficient NLLB-200: Language-specific expert pruning of a massively multilingual machine translation model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3567–3585. Association for Computational Linguistics.

Roman Koshkin, Katsuhito Sudoh, and Satoshi Nakamura. 2024. LLMs are zero-shot context-aware simultaneous translators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1207. 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. ELRA and ICCL.

OpenAI. 2024a. Gpt-4o system card. Accessed: 2025-04-27.

OpenAI. 2024b. Openai o1 system card. Accessed: 2025-04-27.

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.

Daniel Pedersen. 2014. Exploring the concept of transcreation–transcreation as "more than translation". *Cultus: The Journal of intercultural mediation and communication*, 7(1):57–71.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376. Association for Computational Linguistics.

Maja Popović. 2018. *Error Classification and Analysis for Machine Translation Quality Assessment*, pages 129–158. Springer International Publishing, Cham.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

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.

Robert F. Tate. 1954. Correlation between a discrete and a continuous variable. point-biserial correlation. *The Annals of Mathematical Statistics*, 25(3):603–607.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual

NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533. 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. Association for Computational Linguistics.

xAI. 2024. Bringing grok to everyone. Accessed: 2025-04-27.

Shufang Xie, Yingce Xia, Lijun Wu, Yiqing Huang, Yang Fan, and Tao Qin. 2022. End-to-end entity-aware neural machine translation. *Mach. Learn.*, 111(3):1181–1203.

Emmanouil Zaranis, Nuno M Guerreiro, and Andre Martins. 2024. Analyzing context contributions in LLM-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14899–14924. 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. Association for Computational Linguistics.

A Appendix

	BLEU	COMET	M-ETA	Annotator Score
o1	0.39	0.91	0.30	0.52
o1 Mini	0.39	0.91	0.38	0.50
GPT-4o	0.38	0.93	0.26	0.40
GPT-4o Mini	0.34	0.92	0.34	0.30
Claude 3.5 Sonnet	0.22	0.83	0.40	0.22
Claude 3.5 Haiku	0.14	0.80	0.24	0.12
Gemini 1.5 Pro	0.39	0.90	0.40	0.38
Gemini 1.5 Flash	0.29	0.92	0.28	0.34
Grok 2	0.32	0.92	0.36	0.58
DeepkSeek R1	0.00	0.49	0.00	0.00
Llama 3	0.04	0.58	0.04	0.06
MBart05	0.17	0.87	0.10	0.22
NLLB-200	0.22	0.88	0.22	0.14

Table 2: Automatic evaluation results for BLEU, COMET and M-ETA and Human Evaluation Score.
Color key: = Highest BLEU, = Highest COMET, = Highest M-ETA. = Highest Annotator Score.

Popularity Rank	Entity Type	BLEU	COMET	M-ETA
1	Plant	0.3448	0.8354	0.6573
2	Book	0.2245	0.8188	0.2721
3	Person	0.2666	0.8526	0.2704
4	Artwork	0.2096	0.8148	0.2497
5	Food	0.3317	0.8421	0.3646
6	Movie	0.1707	0.8151	0.1812
7	Fictional entity	0.3070	0.8337	0.3685
8	Animal	0.3026	0.8257	0.3846
9	Landmark	0.3026	0.8784	0.2552
10	TV series	0.2078	0.8077	0.1628
11	Place of worship	0.3141	0.8625	0.3070
12	Natural place	0.1638	0.9058	0.6410
13	Musical work	0.3297	0.8558	0.3735
14	Book series	0.2471	0.7992	0.0804

Table 3: Entites ranked by popularity levels along with their average scores.
Color key: = Highest BLEU, = Highest COMET, = Highest M-ETA.

Models	Metric	Entity Types													
		Animal	Artwork	Book	Book Series	Fictional Entity	Food	Landmark	Movie	Musical Work	Natural place	Person	Place of Worship	Plant	TV Series
o1	BLEU	0.3712	0.3236	0.3334	0.3964	0.4659	0.4988	0.4284	0.2953	0.4477	0.2108	0.4243	0.5019	0.7097	0.3202
	Comet	0.9393	0.9018	0.9029	0.8779	0.9384	0.9438	0.9537	0.8965	0.9388	0.9720	0.9426	0.9488	0.9582	0.8862
	M-ETA	0.5000	0.3602	0.3634	0.1940	0.5703	0.5092	0.3413	0.2683	0.3912	1.0000	0.4118	0.4448	0.8182	0.2396
o1 Mini	BLEU	0.3698	0.3204	0.3504	0.3827	0.5013	0.5301	0.4352	0.2143	0.4869	0.2927	0.4034	0.4563	0.5268	0.3186
	Comet	0.9223	0.9063	0.9115	0.8770	0.9359	0.9430	0.9540	0.8991	0.9385	0.9768	0.9410	0.9435	0.9324	0.8838
	M-ETA	0.0000	0.2762	0.3225	0.0448	0.4934	0.4479	0.2857	0.1847	0.5705	1.0000	0.3069	0.3785	0.7273	0.1506
GPT-4o	BLEU	0.5247	0.2926	0.3204	0.3116	0.4544	0.5301	0.4332	0.2423	0.4644	0.2717	0.4211	0.4491	0.6030	0.2835
	Comet	0.9046	0.8938	0.8975	0.8629	0.9122	0.9418	0.9495	0.8858	0.9217	0.9684	0.9276	0.9363	0.9420	0.8801
	M-ETA	0.6667	0.3782	0.4091	0.1493	0.5093	0.4847	0.3651	0.2840	0.5034	0.6667	0.3890	0.4038	0.9091	0.2927
GPT-4o Mini	BLEU	0.4254	0.2755	0.3049	0.3569	0.4230	0.4827	0.4144	0.2517	0.4520	0.3137	0.3789	0.4261	0.5220	0.2968
	Comet	0.8876	0.8782	0.8786	0.8737	0.9090	0.9374	0.9487	0.8943	0.9202	0.9720	0.9213	0.9346	0.9496	0.8845
	M-ETA	0.1667	0.2736	0.2960	0.0597	0.4164	0.4387	0.3016	0.1829	0.3735	0.6667	0.2692	0.3596	0.9091	0.1664
Claude-3.5 Sonnet	BLEU	0.1324	0.1507	0.1685	0.1644	0.2000	0.2349	0.2578	0.1257	0.3372	0.0000	0.1681	0.2243	0.2190	0.1674
	Comet	0.7358	0.8193	0.8329	0.8183	0.7991	0.8123	0.8787	0.8466	0.8939	0.8295	0.8340	0.8463	0.7539	0.8269
	M-ETA	0.8333	0.3448	0.3682	0.0597	0.4934	0.5429	0.3333	0.2526	0.6142	0.6667	0.3688	0.4479	1.0000	0.2439
Claude-3.5 Haiku	BLEU	0.0407	0.1320	0.1377	0.1921	0.1568	0.2241	0.2089	0.1129	0.2528	0.0000	0.1255	0.1492	0.1348	0.1331
	Comet	0.6727	0.7875	0.7851	0.7962	0.7663	0.7913	0.8724	0.7834	0.8557	0.8263	0.8035	0.8346	0.7320	0.7981
	M-ETA	0.8333	0.2787	0.2960	0.1045	0.3979	0.4755	0.2937	0.2073	0.2599	1.0000	0.2948	0.3880	0.8182	0.1535
Gemini-1.5-Pro	BLEU	0.4971	0.3514	0.3689	0.2932	0.4189	0.4678	0.4132	0.2972	0.4914	0.2381	0.3786	0.4714	0.4760	0.2703
	Comet	0.9508	0.8927	0.8920	0.8578	0.9166	0.9371	0.9501	0.8812	0.9371	0.9667	0.9291	0.9412	0.9522	0.8739
	M-ETA	0.5000	0.4871	0.5174	0.1940	0.5570	0.5245	0.3889	0.3345	0.7291	0.0000	0.4616	0.4227	0.8182	0.3286
Gemini-1.5-Flash	BLEU	0.4989	0.2424	0.2656	0.2561	0.3868	0.4001	0.3482	0.2291	0.3242	0.0000	0.3308	0.3846	0.5315	0.2332
	Comet	0.9499	0.8876	0.8932	0.8548	0.9207	0.9292	0.9459	0.8907	0.9248	0.9634	0.9342	0.9393	0.9532	0.8721
	M-ETA	0.8333	0.3045	0.3586	0.0896	0.4695	0.4417	0.3175	0.2439	0.3748	0.3333	0.3513	0.4006	0.9091	0.2023
Grok 2	BLEU	0.3493	0.2808	0.3042	0.3665	0.4617	0.4973	0.4121	0.2571	0.5490	0.3137	0.3912	0.4749	0.4891	0.3171
	Comet	0.9053	0.8923	0.8980	0.8876	0.9250	0.9387	0.9499	0.8902	0.9356	0.9747	0.9358	0.9443	0.9516	0.8885
	M-ETA	0.1667	0.3113	0.3430	0.1343	0.5013	0.4417	0.2460	0.2213	0.5636	1.0000	0.3163	0.3849	0.9091	0.2023
DeepSeek-R1-7B	BLEU	0.0315	0.0076	0.0076	0.0085	0.0034	0.0060	0.0137	0.0049	0.0071	0.0000	0.0044	0.0045	0.0000	0.0090
	Comet	0.5619	0.4891	0.4971	0.4952	0.4752	0.4664	0.4914	0.5036	0.4911	0.5595	0.4858	0.4714	0.4533	0.5038
	M-ETA	0.0000	0.0026	0.0036	0.0000	0.0053	0.0031	0.0079	0.0000	0.0000	0.0000	0.0013	0.0000	0.0000	0.0072
Llama 3	BLEU	0.0128	0.0154	0.0193	0.0407	0.0566	0.0601	0.0642	0.0130	0.0310	0.0000	0.0450	0.0570	0.0000	0.0238
	Comet	0.5322	0.4972	0.4997	0.5121	0.5633	0.5500	0.6507	0.5124	0.5784	0.8461	0.6225	0.6564	0.4895	0.5111
	M-ETA	0.0000	0.0223	0.0409	0.0000	0.1114	0.1258	0.0952	0.0261	0.0328	0.6667	0.0983	0.0946	0.0000	0.0316
mBART-Large-50	BLEU	0.2799	0.1540	0.1573	0.1808	0.1977	0.1655	0.1982	0.0716	0.1282	0.1291	0.1684	0.1949	0.0529	0.1125
	Comet	0.8659	0.8647	0.8692	0.8330	0.8814	0.8848	0.9301	0.8330	0.8897	0.9638	0.8878	0.9007	0.8821	0.8341
	M-ETA	0.0000	0.0823	0.0927	0.0000	0.1088	0.1227	0.1746	0.0488	0.0643	1.0000	0.0888	0.0726	0.2727	0.0488
NLLB-200	BLEU	0.3999	0.1781	0.1805	0.2620	0.2638	0.2141	0.3061	0.1044	0.3141	0.3591	0.2258	0.2895	0.2180	0.2159
	Comet	0.9060	0.8813	0.8870	0.8429	0.8948	0.8719	0.9447	0.8800	0.8998	0.9566	0.9183	0.9146	0.9107	0.8576
	M-ETA	0.5000	0.1244	0.1252	0.0149	0.1565	0.1810	0.1667	0.1010	0.3776	0.3333	0.1575	0.1924	0.4545	0.0488

Table 4: BLEU, COMET, and M-ETA scores for each model and entity type.
Color key: = Highest BLEU, = Highest COMET, = Highest M-ETA.

Type of Error	Definition
Literal Translation	Translation follows the meaning of the source language.
Phonetic Translation	Translation follows how it sounds in the source language.
Word-Level Translation	Translation is done word-for-word from the source language.
Incorrect Entity Name	Used a different or less appropriate entity name.
Incorrect Grammar	Grammar mistakes in the target language.
Incorrect Language	Translated into the wrong language.
Incorrect Formatting	Formatting is wrong, but the translation itself is correct.
Added/Deleted Content	Extra parts added or parts missing compared to the source.
Incorrect Response	Output that doesn't match the source text meaning.
Partial Translation	Only part of the source text is translated.
Romanized Korean	Latin alphabet used instead of proper Korean script.
Gibberish	Output makes no sense at all.

Table 5: Types of translation errors and their definitions.

Models	Metric	Popularity Level				
		Low *142 - 12943*	**Low-mid** *12944 - 29440*	**Mid** *29441 - 62685*	**Mid-high** *62686 - 157350*	**High** *157351 - 6974823*
o1	BLEU	0.3789	0.3764	0.3760	0.3696	0.4377
	COMET	0.9167	0.9173	0.9149	0.9184	0.9297
	M-ETA	0.3171	0.3259	0.3415	0.3579	0.5321
o1 Mini	BLEU	0.3883	0.3791	0.3723	0.3621	0.4180
	COMET	0.9179	0.9200	0.9145	0.9186	0.9300
	M-ETA	0.3028	0.3005	0.3129	0.3013	0.4455
GPT-4o	BLEU	0.3713	0.3689	0.3676	0.3496	0.3910
	COMET	0.9082	0.9122	0.9036	0.9094	0.9109
	M-ETA	0.3618	0.3492	0.4006	0.3589	0.5066
GPT-4o Mini	BLEU	0.3720	0.3615	0.3373	0.3449	0.3607
	COMET	0.9078	0.9082	0.8960	0.9058	0.9059
	M-ETA	0.2530	0.2680	0.2681	0.2841	0.3894
Claude 3.5 Sonnet	BLEU	0.2088	0.2114	0.1996	0.1672	0.1977
	COMET	0.8490	0.8530	0.8374	0.8264	0.8285
	M-ETA	0.3567	0.3777	0.3812	0.3761	0.4995
Claude 3.5 Haiku	BLEU	0.1881	0.1794	0.1557	0.1332	0.1393
	COMET	0.8269	0.8219	0.8032	0.7918	0.7829
	M-ETA	0.2266	0.2538	0.2803	0.2781	0.3914
Gemini 1.5 Pro	BLEU	0.3707	0.3776	0.3743	0.3691	0.4239
	COMET	0.9108	0.9127	0.8995	0.9066	0.9183
	M-ETA	0.4360	0.4589	0.4638	0.4611	0.5994
Gemini 1.5 Flash	BLEU	0.2864	0.2827	0.2948	0.2923	0.3305
	COMET	0.9079	0.9117	0.9035	0.9058	0.9120
	M-ETA	0.2571	0.2964	0.3425	0.3195	0.4444
Grok 2	BLEU	0.4061	0.3779	0.3919	0.3498	0.3869
	COMET	0.9130	0.9157	0.9118	0.9135	0.9173
	M-ETA	0.3018	0.3147	0.3405	0.3488	0.4648
DeepSeek R1	BLEU	0.0091	0.0087	0.0051	0.0052	0.0041
	COMET	0.4955	0.4924	0.4852	0.4837	0.4892
	M-ETA	0.0041	0.0020	0.0020	0.0020	0.0020
Llama 3	BLEU	0.0369	0.0281	0.0376	0.0315	0.0286
	COMET	0.5652	0.5634	0.5513	0.5373	0.5466
	M-ETA	0.0539	0.0315	0.0550	0.0586	0.0765
MBart-50	BLEU	0.1446	0.1411	0.1464	0.1453	0.1442
	COMET	0.8700	0.8742	0.8675	0.8660	0.8735
	M-ETA	0.0640	0.0558	0.0744	0.0809	0.1172
NLLB-200	BLEU	0.2317	0.2399	0.2236	0.2006	0.2042
	COMET	0.8915	0.8933	0.8871	0.8912	0.8884
	M-ETA	0.1667	0.1848	0.1784	0.1547	0.1600

Table 6: BLEU, COMET and M-ETA scores grouped by popularity levels for each model.
Color key: = Highest BLEU, = Highest COMET, = Highest M-ETA.

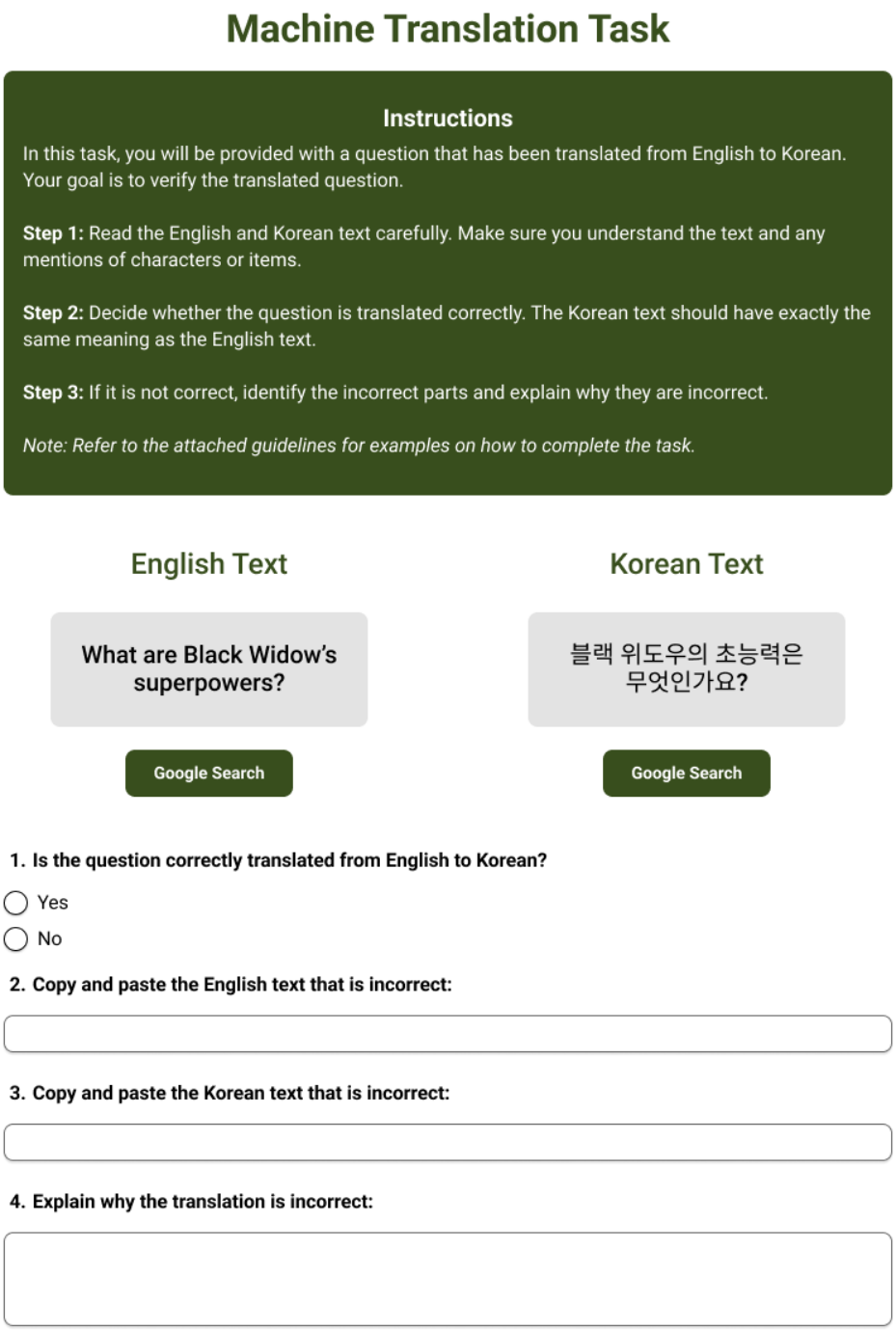

Figure 4: UI used for machine translation human annotation task.