



**The 17th Conference of
The Association for Machine Translation
in the Americas**

www.conference.amtaweb.org

Volume 1: Research Track

Editors:
Eleftheria Briakou, Jeremy Gwinnup, and Shivali Goel
(Research Track Co-chairs); Shivali Goel (Publications
Chair); Jay Marciano (General Conference Chair)

Welcome to the 17th conference of the Association for Machine Translation in the Americas – AMTA 2026!

Dear MT & AI Colleagues and Friends,

Welcome to the Proceedings of AMTA 2026, the 17th biennial conference of the Association for Machine Translation in the Americas. These proceedings bring together research and practical experience from a community that continues to redefine the state of the art in multilingual communication. The contributions collected here represent advances in machine translation, the use of large language models, generative AI, and agentic AI for cross-language use cases, evaluation methodologies, and real-world deployment across an increasingly diverse range of applications.

For this year's conference of the Association for Machine Translation in the Americas, we come together in person in beautiful Québec City, Canada, at the Centre des congrès de Québec from August 31st to September 2nd, and online for our Virtual Day of Tutorials on August 19th.

At our last in-person conference, AMTA 2024, we marked the 30th anniversary of the first AMTA conference, held in Columbia, Maryland, in October 1994, and 70 years since the first public demonstration of machine translation, the famous IBM-Georgetown experiment of 1954. And at this year's conference, we can mark the end of the first decade of Neural Machine Translation, the MT paradigm and architecture that rapidly took over the field and subsequently gave rise to Large Language Models and generative AI.

It is difficult to argue against the ubiquity of MT today. In an April 2026 blog post celebrating 20 years of Google Translate, Google estimated that their technology now translates approximately one trillion words monthly while servicing more than one billion customers. While Google is the recognized leader in the volume of words translated, there are many other heavily used and specialized systems in increasingly automated translation workflows and for new use cases. With such widespread use of MT distributed over so many systems, even approximating the volume of material processed by these systems is a daunting challenge. Add to that the use of LLMs for cross-language use cases, which blurs the lines between data curation, content creation, translation, and quality evaluation, and you must conclude that trying to come up with usage estimates is akin to guessing how many pennies are in a large glass jar on the counter of a country store: it can be nothing more than an educated guess.

While it is unquestionably an indication of acceptable usability, unthinkable high usage does not imply that automated translation is a solved technical problem. Work continues on expanding language support, on improving speech-to-speech systems, on developing systems for interpreting and translating sign language, applying agentic AI to the task of selecting how these technologies can best be applied to a given translation task, and reliable automated assessment of translation quality. With the widespread use of MT and related technologies, we find ourselves in an intense period of developing best practices for the application of these technologies, and, importantly, in exploring how to most effectively measure the quality of their output.

To this end, this year AMTA sponsored an industry working group to propose a methodological framework for conducting meaningful evaluations of translation quality evaluation systems.

Members of this working group, organized by Evelyn Yang Garland and including 23 industry experts, will be presenting a three-hour tutorial on TQE, including a practical introductory guide into both the current landscape of TQE systems and their primary use-cases across the translation industry, a review of primary findings of TQE evaluations from the WMT Shared Task on Automated TQE of recent years, and an overview of the findings and recommendations of an AMTA-sponsored working group, and will be presenting their results in a talk at the in-person conference as well.

What is of growing concern in the age of generative AI, across all industries and use cases, is the authenticity of voices, and it is exactly this concern with authenticity that makes AMTA conferences so valuable. A unique aspect of our conferences is that they bring together users and practitioners from across the MT spectrum of academia, industry, public sector entities, technology providers and, increasingly, individual language professionals, on whose daily work these technologies have had and will continue to have a profound impact. It is this authentic exchange of ideas from person to person, across professional categories, that will allow us to continue to address the challenges that remain.

We are pleased once again with the number and quality of submissions to our conference, which reflect great progress in MT, not only in the scope of supporting ever more languages and in improving and assessing output quality, but also in the use of large language models either as translation systems in and of themselves or, as you will see in many of the papers and presentations this year, in augmenting machine translation systems with additional processing via LLMs.

We are also pleased to award the second AMTA Best Thesis Award, a tradition that we borrowed from our sister organization EAMT. We congratulate this year's winner, Dr. Elizabeth Salesky, for her thesis "Overcoming Vocabulary Challenges in Natural Language Processing," an abstract of which is included in this volume.

As with all our conferences, AMTA 2026 would simply not have been possible without the selfless work of so many people on the AMTA board and organizing committee, all of whom are volunteers. I express my heartfelt thanks, respect, and admiration to each of them. They include:

Marianna Martindale, AMTA Secretary

Alex Yanishevsky, AMTA Vice President, Conference Online Platform

Ed Watts, AMTA Treasurer

Eleftheria Briakou, Peer-review Track and Best Thesis Award organizer

Jeremy Gwinnup, Peer-review Track and Best Thesis Award organizer

Shivali Goel, Peer-review Track, Best Thesis Award organizer, and Publications

Cecilia Yalangozian, Workshops and Tutorials, Presentations Track

Vera Senderowicz Guerra, Workshops and Tutorials, Presentations Track

Georg Kirchner, Presentations Track

Konstantine Boukhvalov, Public Sector Outreach

Derick Fajardo, Communications and Marketing

Lara Shackelford Daly, Sponsorships

Steve Richardson, AMTA Counselor
Alon Lavie, AMTA Consultant
Evelyn Yang Garland, Chair of TQE Workgroup
Kelly Ko, Webmaster

I also express my gratitude to our sponsor, Nations Translation Group, whose support has helped us defray the cost of this conference.

We extend our sincere appreciation to the authors for sharing their research, to the reviewers for their thoughtful evaluations, and, importantly, we thank the participants whose curiosity, collaboration, and commitment continue to advance the field of machine translation and multilingual language technologies.

We hope that these proceedings will serve not only as a record of the current state of the field, but also as a resource that informs future research and inspires continued collaboration across disciplines, institutions, and national boundaries. As machine translation enters its next decade of rapid innovation, the work presented here demonstrates that our community remains well positioned to shape the future of multilingual communication.

Sincerely,
Jay Marciano
AMTA President and AMTA 2026 General Conference Chair

The Second AMTA Best Thesis Award

For the second edition of the AMTA Best Thesis Award, we again received a strong and competitive slate of eligible candidates and theses, resulting in a difficult decision to choose a single Best Thesis Award winner. Each thesis was assessed by multiple reviewers for its relevance to the field of machine translation, the past, current, and potential future impact of the thesis work on the field, and how challenging the thesis's core problem was. To ensure fairness in this single-blind process, reviewers were assigned to avoid institutional or other conflicts of interest, and where necessary, members of the board recused themselves.

A committee appointed by the AMTA Board, consisting of Eleftheria Briakou, Jeremy Gwinnup, and Shivali Goel, oversaw the process, summarised the reviews, deliberated, and ultimately submitted a recommendation for the selected winner to the full AMTA Board for approval.

We are pleased to announce that the 2026 AMTA Best Thesis Award is awarded to:

Dr. Elizabeth Salesky

for her thesis entitled “Overcoming Vocabulary Challenges in Natural Language Processing.”

This thesis takes on the long-standing vocabulary bottleneck in machine translation and other natural language processing applications, exploring more robust and flexible representations of text and their impact on translation quality and cross-lingual generalization. The Proceedings include a short summary of the thesis's main contributions, and we encourage interested readers to explore the full thesis in greater depth. Dr. Salesky will present her work at the 2026 AMTA conference and will receive a prize of \$1000 USD along with a free AMTA membership and waived registration costs.

We thank the following reviewers for their work on this process:

Alon Lavie (Carnegie Mellon University)
Antonios Anastasopoulos (George Mason University)
David Adelani (Mila – Québec AI Institute / McGill University)
Huda Khayrallah (Johns Hopkins University)
Jeremy Gwinnup (Air Force Research Laboratory)
Kevin Duh (Johns Hopkins University)
Lei Li (UC Santa Barbara)
Marine Carpuat (University of Maryland)
Marzena Karpinska (Simon Fraser University)
Rebecca Knowles (National Research Council Canada)
Sweta Agrawal (Instituto de Telecomunicações)
Vered Shwartz (University of British Columbia)

We also thank the European Association for Machine Translation, especially for the long-running Anthony C. Clarke EAMT Best Thesis Award, which served as inspiration for the AMTA Best Thesis Award.

Introduction to the Research Papers Volume

The Association for Machine Translation in the Americas meets in Québec City for its 17th biennial conference. This year's peer-reviewed program reflects a field that is changing quickly, with long-standing questions in machine translation now shaped by large language models and generative AI.

The Research Papers volume of the AMTA 2026 Proceedings includes eleven research papers, along with the abstract for the 2026 AMTA Best Thesis Award. Large language models appear throughout, both as translation systems in their own right and as components added to existing systems. Several papers focus on evaluation and quality estimation. Others address fairness and accessibility, including gender bias, crosslingual disparities, and translation for atypical speech. The volume also includes work on terminology, post-editing in production and government settings, parallel data alignment, and the use of large language models in specialized domains. As is usual for AMTA, the papers come from academia, industry, and government.

Many of these topics also appear in the Presentations track, and we expect the two programs to complement each other and to give a broad view of where MT research stands and where it is going.

We thank the reviewers and emergency reviewers, the AMTA organizing committee, and the past research chairs for their help in putting this program together.

Sincerely,

Eleftheria Briakou, Jeremy Gwinnup, and Shivali Goel
Research Track Co-Chairs, 2026

Program Committee (Peer-Reviewed Track)

Alon Lavie (Carnegie Mellon University)

Ankur Bhatnagar

Cecilia Yalangozian (ModelFront)

Chutong Meng (George Mason University)

David Guzmán (Mila – Québec AI Institute)

Eleftheria Briakou (Google)

Elizabeth Nielsen (Google)

HyoJung Han (Google)

Jay Marciano (AMTA)

Jeremy Gwinnup (Air Force Research Laboratory)

Konstantine Boukhvalov (ManpowerGroup)

Marianna Martindale (University of Maryland)

Nathaniel Robinson (Johns Hopkins University)

Shivali Goel (Adobe)

Stephen Richardson (Brigham Young University)

Sweta Agrawal (Instituto de Telecomunicações)

Vera Senderowicz Guerra (Welocalize)

Content

1-2	AMTA Best Thesis Award Abstract: Overcoming Vocabulary Challenges in Natural Language Processing
3-27	Translation-CoT: A Human-Inspired Chain-of-Thought Framework for Multilingual LLM Translation
28-41	Toward Equitable Machine Translation for Atypical Speech: An LLM Post-Correction Approach
42-79	LLMs as Translator Training Partners: A Multi-Agent Approach
80-89	Predict and Fix: A Unified Model for Translation Quality Estimation and Post-Editing
90-100	BRIDGE-MT: A Benchmark for Role Interactions and Dependencies in Machine Translation Gender Evaluation
101-134	Improving Term Evaluation in Machine Translation: Variation Matters
135-145	Layout-Based Chunk Alignment: Utilizing Visual Information to Collect Parallel Texts From Image Documents
146-158	Crosslingual Disparities in LLM Performance: Challenges for MT as Mitigation
159-173	Translators' Perceptions and Edit Traces: Quality of MT as a Tool in Canadian Parliamentary Translation
174-185	Document Summarization for AI-based Post-Editing
186-201	Benchmarking Large Language Models for Game Localization Quality Assurance: A Cross-Model, Cross-Lingual Analysis

Overcoming Vocabulary Challenges in Natural Language Processing

Elizabeth Salesky

Johns Hopkins University, Baltimore, MD, USA

elizabeth.salesky@jhu.edu

Abstract

Natural language processing (NLP) applications such as machine translation (MT) have the potential to overcome language barriers and reduce the digital divide, improving information access for speakers of all languages. Despite recent advances it remains challenging to build language models which generalize across the world’s 7,000 languages and 150 scripts—a challenge which begins from one of the very first steps, designing the model vocabulary. Language models are typically defined over a finite set of inputs, even if designed to have an “open vocabulary” through common tokenization techniques like subword segmentation. Current tokenization approaches are primarily designed around how computers handle text storage (using bytes and Unicode) rather than how languages and scripts are structured, which limits their effectiveness and adaptability across settings. These challenges are exacerbated in multilingual models where balancing computational costs and coverage results in a “**vocabulary bottleneck**.”

This thesis focuses on techniques to build adaptive and robust open-vocabulary language models. We address three core challenges pertaining to model vocabularies within the context of machine translation: optimizing vocabulary size, enhancing model robustness, and overcoming the vocabulary bottleneck in multilingual models. We first demonstrate the importance of tuning the size of subword model vocabularies and propose a method to do so more efficiently through online incremental vocabulary expansion. We then propose a line of work in which we overcome the vocabulary bottleneck by replacing the embedding matrix with representations built on visually rendered text (pixels). Our vocabulary-free approach improves model robustness across a linguistically diverse set of languages and also improves cross-lingual transfer both within and across scripts, particularly for under-resourced languages. Finally, we present a new benchmark for a real-world task, visually-situated translation of text in natural images, where direct multimodal models can reduce error propagation and improve robustness compared to cascaded approaches.

1 Summary

Machine translation (MT) is a branch within natural language processing (NLP) focused on using computational methods to translate text or speech from one language into another. Machine translation has long served as a testing ground for new methods in NLP due to its extensive history within the field, the availability of data for benchmarking new approaches, and its direct relevance to practical applications. While the quality of MT has advanced significantly since neural models were introduced (Luong and Manning, 2015; Kocmi et al., 2023), one critical but often overlooked component that has seen little change since 2015 is **tokenization**. MT models, like other NLP models, rely on predefined model vocabularies often created through data-driven segmentation to support an “open vocabulary.” The open-vocabulary challenge refers to effective handling of any input words or phrases, even if they are not present in model’s training data or fixed vocabulary. Current solutions, while generally effective, are primarily designed around how computers handle text storage (using bytes and Unicode), rather than how languages and scripts are linguistically structured, which can limit their effectiveness across different languages and settings.

In this thesis, we focus on overcoming three vocabulary challenges in machine translation; namely, optimizing vocabulary size, models’ robustness to noise, and the vocabulary bottleneck in multilingual models.

Constructing a model vocabulary involves a delicate balance between enabling sufficient expressivity and robustness, and managing computational costs. Optimizing vocabulary sizes is a challenging and expensive task (Ding et al., 2019), particularly in the absence of reliable proxy metrics that correlate strongly with downstream tasks—necessitating full training runs to extrinsically compare vocabulary effectiveness for a downstream performance, which is computationally intensive and so under-explored. We demonstrate the importance of optimizing model vocabularies, and present the first work on expanding neural model vocabularies to address this challenge, enabling efficient online vocabulary optimization (Salesky et al, 2018).

Data-driven techniques to create open vocabularies potentially allow for an open vocabulary, where any previously unobserved word can be represented as a combination of more frequently observed substrings or characters. However, common training corpora can lack sufficient coverage to create a generalizable vocabulary, particularly evident in morphologically-rich languages where only a few possible linguistic variations may be observed, or languages with non-standardized orthographies where variations in spelling and encoding further contribute to data sparsity. No matter the size of the vocabulary, there will be unseen input at inference time due to common phenomena in the wild such as spelling errors or unexpected codepoints, to which machine translation models notoriously lack robustness (Belinkov and Bisk, 2018). Previous methodologies do not adequately address these challenges, resulting in disparities in performance across languages and settings. We introduce **visual representations of text** for truly open-vocabulary tokenization (Salesky et al, 2021). We demonstrate that translation directly from pixels substantially improves robustness to a variety of forms of noise and results in more compositional text representations. Moreover, they eliminate the need to optimize a discrete vocabulary.

In multilingual settings, balancing capacity and coverage across diverse languages within a fixed parameter budget creates a **vocabulary bottleneck**. As we expand the number of languages to be covered by a particular model, with a finite vocabulary we inherently must reduce the capacity allocated to each language and script; on the other hand, if we maintain capacity per language, the number of parameters allocated to the vocabulary quickly explode and the model becomes intractable to train. A wide body of work has shown that multilingual models are affected by a language’s script (Muller et al., 2021; Rust et al., 2021; Pfeiffer et al., 2021), necessitating vocabulary expansion and adaptation techniques. Recent work which expands the LLaMA-2 (Touvron et al., 2023) model vocabulary from English-only to 534 languages found that increasing the vocabulary size from 32k to 260k alone resulted in a 30% parameter increase (Lin et al., 2024). Another recent work found that a 4× increase in vocabulary size caused masked language model pretraining to take 2.5× longer (Liang et al., 2023). We identify the architectural requirements to apply visual representations of text to multilingual settings and at scale. We demonstrate that our approach enables **native cross-lingual transfer between scripts** with significantly greater positive transfer, particularly for low-resource languages (Salesky et al, 2023), and comes without the computational cost of byte-level or larger subword vocabularies.

Finally, we turn to a real-world visually-situated task, translation of text in natural images. While this task is typically approached with a cascade of multiple separate models including OCR and MT, we create a taxonomy of typical OCR errors and their impact on downstream translation performance. We present a novel benchmark that allows the evaluation of direct architectures such as those developed in the previous two chapters on this task, and show that multimodal models translating directly from pixels can reduce error propagation and improve robustness for this capability (Salesky et al, 2024).

This thesis advances robust open-vocabulary translation by introducing approaches such as vocabulary expansion and vocabulary-free, pixel-based text representations that bypass traditional tokenization bottlenecks. Furthermore, it presents a novel benchmark for visually-situated translation of text in natural images and demonstrates that end-to-end multimodal models can perform more robustly on this task. Together, these contributions provide highly adaptable, scalable models that improve translation quality and cross-lingual transfer across low-resource languages, diverse scripts, and real-world tasks.

Translation-CoT: A Human-Inspired Chain-of-Thought Framework for Multilingual LLM Translation

Tabia Tanzin Prama

[tabia.tanzin.prama@uvm.edu]

Computational Story Lab, Vermont Complex Systems Institute,
Departments of Computer Scienc, University of Vermont, Burlington, VT 05405, USA

Juniper L. Lovato

[Juniper.Lovato@uvm.edu]

Computational Ethics Lab, Vermont Complex Systems Institute,
Department of Computer Science, University of Vermont, Burlington, VT 05405, USA,
Complexity Science Hub, Vienna, Austria

Christopher M. Danforth

[Chris.Danforth@uvm.edu]

Computational Story Lab,
Vermont Complex Systems Institute, Vermont Advanced Computing Center,
Department of Mathematics and Statistics, University of Vermont, Burlington, VT 05405, USA

Peter Sheridan Dodds

[Peter.Dodds@uvm.edu]

Computational Story Lab,
Vermont Complex Systems Institute, Vermont Advanced Computing Center,
Departments of Computer Science, University of Vermont, Burlington, VT 05405, USA;
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA;
Complexity Science Hub, Metternichgasse 8, 1030 Vienna, Austria

Abstract

Large language models (LLMs) have transformed machine translation, yet mistranslations, hallucinations, and unnatural phrasing still limit their effectiveness, particularly for low-resource languages. We propose Translation-CoT, a chain-of-thought prompting strategy that breaks translation into structured stages (lexical retrieval, grammatical analysis, and topic identification), followed by a refinement step to improve fluency, tone, and idiomatic expression. We evaluate Translation-CoT across 14 languages from 14 language families and multiple LLMs (GPT-4o, GPT-4o-mini, LLaMA 3.1, and Gemma 2), with GPT-4o performing best overall, in both English $\leftrightarrow$ non-English (X) translation settings. Compared with zero-shot prompting, in-context learning, and existing chain-of-thought prompting methods (Tree-of-Thought (ToT) and Learning-Oriented Prompting (LOT)), Translation-CoT outperforms these prompting strategies on multilingual machine translation across BLEU, ChrF, and METEOR, with especially strong gains in the more difficult English $\rightarrow$ non-English (X) setting and in low-resource languages. Human evaluation further shows higher preference scores and lower Multidimensional Quality Metrics (MQM) penalty scores, indicating fewer mistranslations, omissions, awkward phrasing, and hallucinations with Translation-CoT. Overall, our results show that structured, task-aware prompting is an effective approach for improving multilingual translation quality and robustness in LLMs.

1 Introduction

Machine Translation (MT) is the study of using computers to translate text from one language into an-

other. The concept of MT was first proposed in 1947, formalized in a 1949 memorandum [Weaver \(1955\)](#), just two years after the development of the first computer, the Electronic Numerical Integrator and Com-

puter. Since then, MT has been recognized as one of the most challenging tasks in natural language processing (NLP). Early MT systems were primarily developed for military purposes. In 1954, Georgetown University, in collaboration with International Business Machines (IBM) Corporation, conducted the first Russian–English MT experiment using the IBM-701 computer, marking the realization of the MT dream. Over time, with the availability of bilingual corpora, corpus-based approaches became dominant, particularly after the 2000s Wang et al. (2021). Three key corpus-based MT paradigms emerged: Example-based machine translation (EBMT), statistical machine translation (SMT), and neural machine translation (NMT), with NMT experiencing significant advances in recent years Zhang and Zong (2020).

Conventional NMT models are trained in a supervised manner and require large-scale parallel corpora Liu et al. (2020); Raffel et al. (2019); Xue et al. (2021). This makes them less effective for low-resource languages (LRLs), which lack sufficient parallel data Prama and Anwar (2025). While researchers have made progress in extracting Schwenk et al. (2019) and annotating Goyal et al. (2021); team et al. (2022) parallel datasets for LRLs, such efforts demand substantial human labor. To address these challenges, the research community has explored unsupervised MT methods Ren et al. (2020) that operate without parallel corpora. LLMs have recently shown remarkable potential in NLP Yang et al. (2024); Dubey et al. (2024); OpenAI et al. (2023) and have been applied to NMT tasks. Studies Robinson et al. (2023); Zhu et al. (2023b) indicate that while LLMs excel in translating high-resource languages, they struggle with LRLs Joulin et al. (2016); team et al. (2022), where data is often scarce and difficult to obtain Prama et al. (2025a). Compared to traditional NMT, LLMs offer several qualitative advantages. They allow controllability of style and language variety through prompting and in-context learning Brown et al. (2020); García et al. (2023); Agrawal et al. (2022), exhibit inherent document-level translation capabilities Wang et al. (2023a); Karpinska and Iyyer (2023), produce less literal translations Raunak et al. (2023), and handle challenging linguistic phenomena such as idioms and ambiguous expressions more effectively. As a result, LLMs are increasingly surpassing NMT models in versatility Peng et al. (2023); Hendy et al. (2023).

Recent research has leveraged in-context learning (ICL) Brown et al. (2020); Dong et al. (2022) to enable LLMs to perform translation without parameter updates, as well as supervised fine-tuning with parallel data Li et al. (2023); Chen et al. (2021); Alves et al. (2023). However, LLM training still demands vast multilingual corpora, and the imbalance in language distribution leads to poor performance for many LRLs Jiao et al. (2023); Hendy et al. (2023); Oni and Prama (2025). While prior work has demonstrated impressive results in high-resource settings such as English–German translation Vilar et al. (2022), the multilingual translation ability of LLMs remains underexplored. In this work, we follow the ICL paradigm to study LLMs for multilingual machine translation (MMT), focusing on two research questions:

RQ1: How do LLMs perform MMT across a large set of languages? We evaluate English-centric LLMs such as LLaMA-3.1 Dubey et al. (2024) and multilingual models including Gemma 2, GPT-4o-mini, and GPT-4o OpenAI et al. (2023) across 14 languages from 14 distinct language families and 28 English-centric translation directions. Appendix A provides the full list of languages and their corresponding families. Results indicate that multilingual translation capabilities of LLMs continue to improve, with GPT-4o achieving the strongest performance among the evaluated models. Notably, GPT-4o outperforms the supervised MMT system NLLB team et al. (2022) for several higher-resource languages, including Estonian, Korean, Bangla, Filipino, and Azerbaijani. However, it does not surpass commercial systems such as Google Translate, which consistently achieves higher performance across all languages.

RQ2: How can we improve LLM translation performance? We propose Translation-CoT, a chain-of-thought prompting strategy Wei et al. (2022) inspired by professional translation workflows Baker (1992); Koehn (2009); Bowker (2002); Hatim and Munday (2005). Unlike generic CoT-based MT methods Nguyen and Xu (2025); Dai et al. (2025), Translation-CoT explicitly guides LLMs to analyze lexical, grammatical, and contextual information before generation, followed by a refinement step for fluency and correctness. We evaluate Translation-

CoT on 14 languages in both English $\leftrightarrow$ non-English directions using LLaMA-3.1, Gemma 2, GPT-4o-mini, and GPT-4o. Across BLEU, ChrF, and METEOR, Translation-CoT consistently improves over zero-shot, ICL, and CoT-style baselines, reducing mistranslations, omissions, and stylistic errors. Gains are strongest for Eng $\rightarrow$ X, while X $\rightarrow$ Eng also shows consistent improvement, especially in METEOR. Compared with external systems, Translation-CoT outperforms No Language Left Behind (NLLB) in several higher-resource languages, but remains below Google Translate in BLEU. Unlike dictionary-based prompting Ghazvininejad et al. (2023) or ICL with curated examples Agrawal et al. (2022), Translation-CoT requires no external resources or domain-specific preparation. Human preference and MQM evaluations Lommel (2013) further confirm that it produces more accurate and fluent translations with fewer errors.

2 Our Translation-CoT Framework

We introduce the Translation-CoT framework. As depicted in Figure 1, Translation-CoT consists of three steps: Preparation Phase, Solution Phase, and Final Translation.

Step 1: Preparation Phase Similar to the initial understanding and interpretation phase that human translators undertake Gile (1995), the preparation phase requires the LLM to first analyze the source text and elicit three aspects of knowledge generally beneficial to translation, though the generated knowledge may contain noise He et al. (2023). Lexicon¹ is crucial for conveying the intended meaning and ensuring faithfulness Baker (1992). Additionally, identifying and maintaining a list of keywords ensures consistent translation of specific terms across different parts of the text. Several studies have emphasized the importance of grammar in translation. Grammar functions as the structural backbone of language Baker (1992); Newmark (1988). Misunderstandings arising from poor grammar comprehension can have significant consequences, particularly in technical translations where precision is crucial Iroda and Risolat (2025).

Step 2: Solution Phase In this phase, the lexical, grammatical, and topical knowledge gathered during the preparation stage is used to produce an

initial translation in the target context (Figure 1, Step 2). The focus here is on refining fluency and tone so the translation reads as if it were originally written in the target language, preserving the author’s intent and engaging the audience Pfau (2015); Bassnett and Lefevre (1990); Taia.io (2025). Awkward or overly literal translations can easily weaken clarity and credibility, so this step also addresses idioms and ambiguities that have no direct equivalent, replacing them with culturally appropriate, natural-sounding expressions Bassnett and Lefevre (1990); Pfau (2015). By doing so, the translation maintains its meaning, avoids misunderstandings, and respects the cultural nuances embedded in the original text. Step 2 in Figure 8 illustrates this solution-phase reasoning.

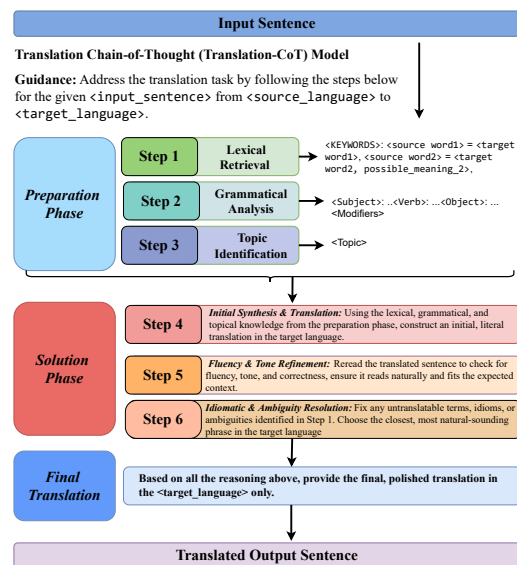


Figure 1: Architecture of the Translation-CoT framework. The framework translates source sentences into target languages in three phases: preparation, solution, and final translation. In the preparation phase, it identifies lexicons, grammatical structure, and topics in the source sentence. The solution phase generates the translation, followed by post-selection refinement to improve fluency, adjust tone, and resolve idiomatic expressions and ambiguities.

Step 3: Final Translation In the final step of the Translation-CoT framework, the LLM generates

¹Lexicons are essential words or phrases that convey the core meaning of a text and serve as focal points for understanding the main idea

the final translated output. Appendix C.3 shows the Translation-CoT prompt framework.

3 Experiments

Dataset. We evaluate multilingual translation performance using the FLORES-101 dataset Goyal et al. (2021)². Our study focuses on 28 translation directions across 14 language families: Indo-European, Mongolic, Atlantic-Congo, Afro-Asiatic, Turkic, Dravidian, Kra-Dai, Austro-Asiatic, Sino-Tibetan, Austronesian, Uralic, Kartvelian, Japonic, Koreanic, and Nilo-Saharan. For each family, we select one representative language, enabling a balanced evaluation across typologically diverse groups. Translation is performed in both directions (English $\Leftrightarrow$ language "X"), capturing asymmetries in model performance between source and target roles. The selected languages and their corresponding families are listed in Appendix A.

Models. We evaluate four widely used language models with diverse capabilities. Llama 3.1 8B Instruct Dubey et al. (2024) and Gemma 2 9B IT Riviere et al. (2024) are open-source models with efficient architectures (e.g., grouped-query attention) and billion-scale parameters. GPT-4o-mini OpenAI et al. (2023) is a cost-efficient model with a 128K-token context window and multimodal support, while GPT-4o OpenAI et al. (2023) is a more advanced multimodal model designed for high-performance and creative tasks. For all translation experiments, we used the API with a deterministic setting of temperature = 0, following prior work Peng et al. (2023) showing that this configuration improves translation quality. We set max tokens = 500 to allow sufficient output length across languages. Since decoding was deterministic, we ran each translation once per prompt-language pair rather than averaging over multiple runs, which also made large-scale evaluation more computationally efficient.

Metric. We evaluate LLM performance using BLEU (Bilingual Evaluation Understudy) Papineni et al. (2002) and ChrF (Character-level F-score) Popovic (2015), which allow for comprehensive assessment across languages. Additionally, we employ METEOR (Metric for Evaluation of Translation with Explicit Ordering) Banerjee and Lavie (2005), which

addresses the evaluation of semantic equivalence limitations of BLEU by incorporating stemming and synonymy matching. Corpus-level BLEU and ChrF scores were computed using SacreBLEU version 2.6.0 (Post, 2018), with one reference translation per source sentence. The corresponding evaluation signatures were BLEU: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.6.0 and ChrF: nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.6.0. METEOR scores were computed using the NLTK implementation by averaging sentence-level scores, with the references and hypotheses tokenized using whitespace splitting. Together, BLEU, ChrF, and METEOR provide a multidimensional evaluation of machine translation quality.

Comparative Methods. For comparison, we evaluate Translation-CoT against several prompting baselines. **Zero-shot** uses a direct translation prompt with temperature set to 0. **In-context learning (ICL)** provides five translation exemplars in the prompt, following prior work showing that exemplar-based prompting can improve MT performance and that both exemplar selection and number affect results Hendy et al. (2023); Agrawal et al. (2022); Zhu et al. (2023b). We also compare against two existing CoT-based prompting strategies: Tree-of-Thought (ToT) Nguyen and Xu (2025), which uses general reasoning over candidate translations, and Learning-Oriented prompting (LOT) Dai et al. (2025), which incorporates structured linguistic guidance. The full prompts for all methods are provided in Appendix C.

4 Results and Discussion

RQ1: Benchmarking LLMs for Multilingual Machine Translation We evaluate the multilingual translation capabilities of LLMs in both directions, X $\Leftrightarrow$ English. Figure 2 reports BLEU scores for each language alongside Google Translate and NLLB. All multilingual LLMs show meaningful translation ability, suggesting that cross-lingual alignment emerges even from largely unsupervised pretraining. Performance generally improves with model scale and recency: GPT-4o consistently outperforms GPT-4o-mini and surpasses LLaMA-3.1 and Gemma-2 across most languages. Despite these gains, specialized MT

²In the benchmarking experiment, we evaluate LLMs on the first 200 sentences of each translation direction’s test set, accounting for the prohibitive API (Application Programming Interface) cost of evaluating a large number of languages. In the analysis experiment, we use the full test set.

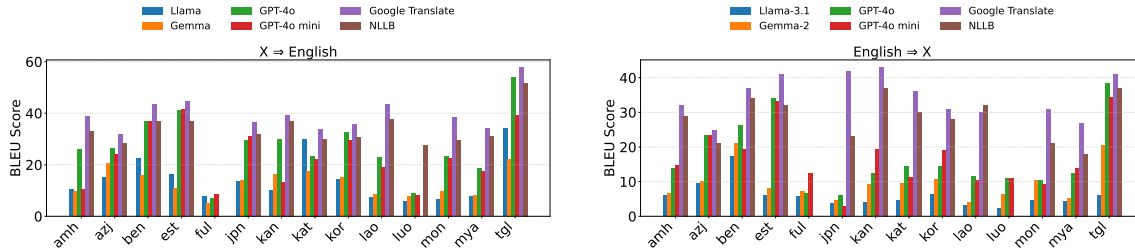


Figure 2: Translation performance (BLEU) of GPT-4o, GPT-4o-mini, Llama 3.1, and Gemma 2 across the evaluated languages. “ $X \rightarrow \text{Eng}$ ” and “ $\text{Eng} \rightarrow X$ ” denote translation into English and from English, respectively. BLEU scores are averaged over the three prompt strategies used in this experiment for each language.

systems remain stronger. Google Translate achieves the highest BLEU scores for nearly all languages, and NLLB [team et al. \(2022\)](#) also outperforms most LLMs. GPT-4o approaches these baselines for some higher-resource languages, such as Bangla, Estonian, and Filipino, but still lags overall, especially for low-resource languages. Among LLMs, GPT-4o obtains the highest average BLEU, ChrF, and METEOR scores (Appendix D). Translation performance is also uneven across directions and languages. Across models, $X \rightarrow \text{English}$ is consistently stronger than $\text{English} \rightarrow X$, reflecting an English-centric bias. Results are generally better for Indo-European languages and weaker for typologically distant or low-resource languages such as Fula and Luo, likely due to both English-dominant pretraining data and structural linguistic differences [Terry et al. \(2024\)](#); [Zhu et al. \(2023b\)](#). Overall, modern LLMs show strong and improving multilingual translation ability, but they have not yet surpassed specialized systems such as Google Translate and NLLB.

RQ2: Enhance LLM’s Translation Performance by Translation-CoT

For consistency, we compare prompting strategies within each LLM. As shown in Figure 3, Translation-CoT is broadly effective and generally improves performance across most languages in both translation directions. In the $X \rightarrow \text{English}$ setting (Appendix E, Table 3), Translation-CoT consistently outperforms Zero-shot and In-Context Learning across models. GPT-4o shows clear BLEU gains for languages such as Bangla, Azerbaijani, Estonian, and Lao; GPT-4o-mini improves in most cases; LLaMA-3.1 shows substantial gains for Filipino, Estonian, and Georgian; and Gemma-2 shows

moderate but consistent improvements, particularly for mid-resource languages. Overall, Translation-CoT matches or exceeds the strongest baseline for most languages, although gains are smaller for low-resource languages such as Fula and Luo.

In the $\text{English} \rightarrow X$ direction (Appendix E, Table 4), improvements are more mixed but remain consistent. Translation-CoT generally outperforms Zero-shot and remains competitive with In-Context Learning, with noticeable gains for GPT-4o and GPT-4o-mini in languages such as Filipino, Georgian, and Mongolian. LLaMA-3.1 and Gemma-2 show smaller but stable improvements. However, In-Context Learning performs comparably or slightly better for some languages and metrics, suggesting that prompt effectiveness is more sensitive when translating into non-English languages. Overall, these results indicate that Translation-CoT improves translation quality by incorporating lexical, grammatical, and topical reasoning before generation, with the clearest gains in BLEU and ChrF and more modest gains in METEOR.

Comparison of CoT-based Prompting Strategies. We compare Translation-CoT with two reasoning-based prompting baselines: Tree-of-Thought (ToT) [Nguyen and Xu \(2025\)](#), which relies on general-purpose reasoning, and Learning-Oriented prompting (LOT) [Dai et al. \(2025\)](#), which incorporates structured linguistic supervision. Unlike these approaches, Translation-CoT explicitly follows a translation-oriented reasoning process: it extracts lexical, contextual, and grammatical information before generation, and then applies a refinement step to improve fluency and naturalness. Figure 4 reports results across 14 languages from diverse language

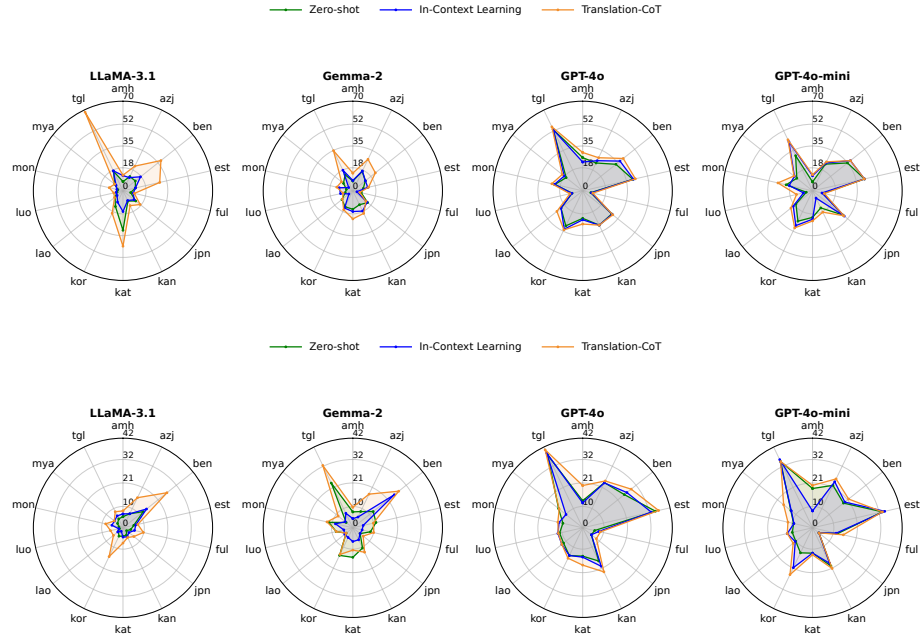


Figure 3: Multilingual translation performance (BLEU) of selected large language models (LLMs) for non-English $\rightarrow$ English (top) and English $\rightarrow$ non-English (bottom) translation across three prompt strategies: Zero-shot, In-context Learning, and Translation-CoT. In almost all cases (8 out of 14 languages), the Translation-CoT prompt outperforms other prompt strategies in BLEU score for LLM performance.

families using GPT-4o, the strongest model in our study.

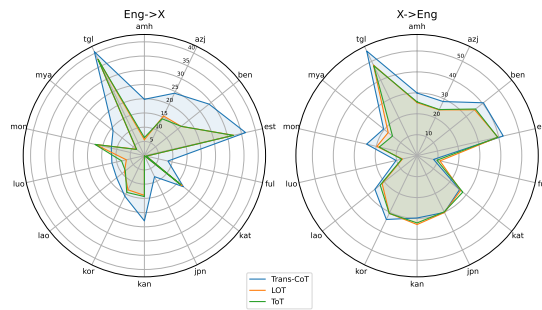


Figure 4: Comparison of CoT-based prompting strategies (Translation-CoT, ToT, and LOT) for multilingual translation performance (BLEU scores) across 14 languages from diverse language families using GPT-4o, for both translation directions (left: Eng $\rightarrow$ X, right: X $\rightarrow$ Eng).

For X $\rightarrow$ Eng, Translation-CoT achieves the

highest BLEU score for 10 of the 14 languages, with the largest gains observed for Filipino, Mongolian, Bangla, and Azerbaijani. Its performance is comparable to the strongest baseline for Japanese, whereas ToT or LOT performs better for Fula, Georgian, and Kannada. Translation-CoT also achieves the highest METEOR score across all languages, although the ChrF results are mixed. For Eng $\rightarrow$ X, Translation-CoT demonstrates a stronger and more consistent advantage, achieving the highest BLEU score for 13 of the 14 languages. The largest improvements are observed for Amharic, Bangla, Burmese, and Azerbaijani, with Mongolian as the only exception. Translation-CoT again achieves the highest METEOR score across all languages, while outperforming the baselines in ChrF for seven languages. Overall, Translation-CoT provides the most consistent improvements in BLEU and METEOR, particularly for Eng $\rightarrow$ X translation. However, the mixed ChrF results indicate that the effectiveness of translation-oriented reasoning varies across languages and evaluation metrics (See Appendix F Figure 13).

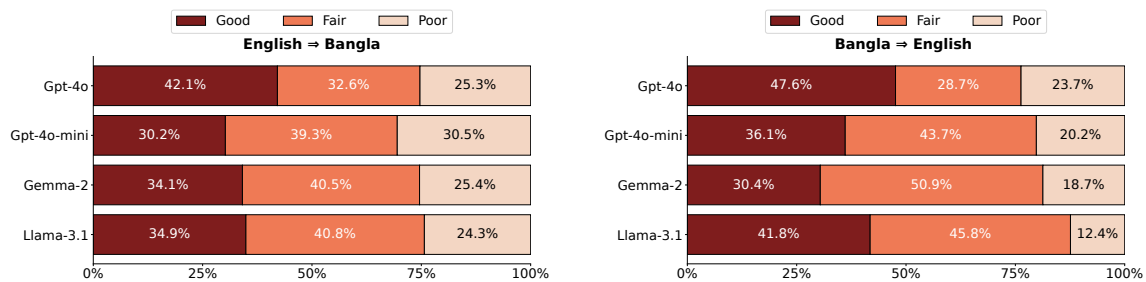


Figure 5: Human preference study comparing Translation-CoT with Zero-Shot and In-Context Learning for LLMs (GPT-4o, GPT-4o-mini, LLaMA 3.1, and Gemma 2).

We further compare Translation-CoT with Google Translate and NLLB across both translation directions (Appendix F, Table 5). In the $X \rightarrow \text{Eng}$ setting, Translation-CoT outperforms NLLB in five languages: Estonian, Korean, Bangla, Filipino, and Azerbaijani. In the $\text{Eng} \rightarrow X$ direction, it surpasses NLLB in Azerbaijani, Estonian, and Filipino, although the improvements are smaller. However, Google Translate remains strongest across nearly all languages, showing that a gap remains between prompted LLM translation and specialized MT systems. Overall, Translation-CoT provides consistent improvements over existing CoT-style prompting methods, improves semantic alignment, and is most competitive in higher-resource settings while still lagging for low-resource languages.

Human Evaluation. We conducted a human preference study on the English $\Leftrightarrow$ Bangla translation task using 200 samples. The evaluation was performed independently by three graduate students from Bangladesh who were fluent in both Bangla and English. Participation was voluntary, and the annotators did not receive financial compensation. For each test sample, annotators were presented with the source sentence and three translations generated using Zero-Shot, In-Context Learning, and Translation-CoT prompting. They rated each translation as *Good*, *Fair*, or *Poor* based on its accuracy, fluency, adequacy, and overall translation quality. Inter-annotator agreement was measured using Fleiss’ κ Fleiss (1971), yielding a score of $\kappa = 0.78$, which indicates substantial agreement among the three annotators. The complete annotation instructions and rating criteria are provided in Appendix G. Figure 5 shows Translation-CoT received more favorable human ratings, indicating their effectiveness in pro-

ducing higher-quality translations compared to Zero-Shot and In-Context Learning.

MQM Evaluation. To identify the specific aspects of translation quality improved by Translation-CoT, we conducted a Multidimensional Quality Metrics (MQM) evaluation (Lommel, 2013) using the same 200 samples from the English $\Leftrightarrow$ Bangla test sets. Following prior MQM-based evaluation procedures (Freitag et al., 2021; He et al., 2023), the annotators independently examined each translation, identified the corresponding error spans, assigned each error to one of four predefined categories (*mistranslation*, *awkward style*, *untranslated text*, or *omission*) and labeled its severity as either minor or major. Each identified error received a penalty based on its severity, and the penalties were summed to obtain an overall MQM score for each translation. Lower MQM scores therefore indicate fewer or less severe translation errors. The annotators were provided with standardized instructions describing the error categories, severity definitions, and penalty weights; the complete guidelines are provided in Appendix G. As shown in Table 1, Translation-CoT significantly outperforms both Zero-Shot and In-Context Learning prompting strategies in terms of MQM scores. In terms of error categories, the improvements achieved by Translation-CoT are primarily due to reductions in mistranslations, awkward style, untranslated text, and omission errors, as illustrated in Figure 6.

LLM Hallucinations. Hallucination in natural language generation (NLG) refers to the production of content that is nonsensical or unfaithful to the given source text Filippova (2020), and remains a persistent challenge for large language models (LLMs) Zhang et al. (2023).

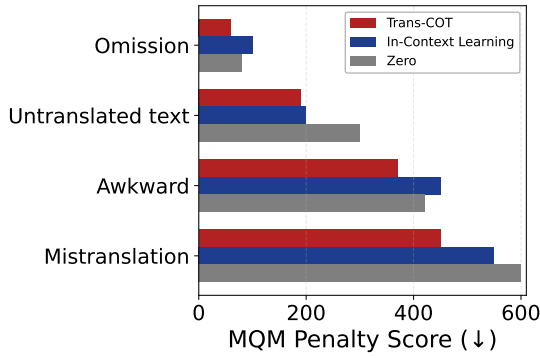


Figure 6: MQM penalty scores across different error categories for 200 test sentences from each of the English↔Bangla test sets. Lower scores indicate less severe translation errors.

To examine this issue, we conducted a human evaluation of hallucination errors across three prompt strategies: Zero-Shot, In-Context Learning, and Translation-CoT. Using the same set of 200 sampled sentences from each of the English↔Bangla test sets, we evaluated four LLMs: LLaMA-3.1, Gemma-2, GPT-4o, and GPT-4o-mini. Following the definition in [Guerreiro et al. \(2023\)](#), translators annotated each translation to indicate the presence or absence of hallucination errors. Following [Guerreiro et al. \(2023\)](#), the same three bilingual annotators independently labeled each translation for the presence or absence of hallucination, with final labels determined by majority vote. Inter-annotator agreement was measured using Fleiss’ κ ($\kappa = 0.86$), and the annotation criteria are provided in Appendix G.2.

As shown in Figure 7, Translation-CoT consistently outperforms both Zero-Shot and In-Context Learning in reducing hallucinations. We attribute this improvement to a fundamental difference between Translation-CoT and In-Context Learning: Whereas In-Context Learning relies solely on example-based prompting, Translation-CoT explicitly incorporates translation-related knowledge into the prompt, which influences the probability distribution over the next token. This additional guidance enables the model to redistribute token probabilities away from erroneous or hallucinated outputs, thereby enhancing translation faithfulness and overall quality.

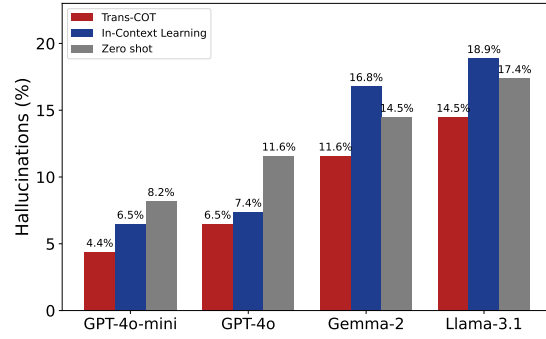


Figure 7: Ratio of hallucinations in generated translations for 200 test sentences from each of the English↔Bangla test sets. Human annotators labeled each output as either containing or not containing a hallucination error.

5 Related Work

Building a separate bilingual translation system for every language pair becomes infeasible as the number of supported languages grows. Multilingual machine translation (MMT) offers a scalable alternative [Johnson et al. \(2016\)](#), yet achieving high quality and efficiency remains challenging [team et al. \(2022\)](#).

Recent advances in LLMs [OpenAI et al. \(2023\)](#); [Touvron et al. \(2023\)](#) have shown strong performance in neural machine translation (NMT) [Jiao et al. \(2023\)](#); [Hendy et al. \(2023\)](#). Two main strategies have emerged: in-context learning (ICL) and fine-tuning. LLMs can perform new tasks from in-context exemplars and templates [Brown et al. \(2020\)](#); [Prana et al. \(2025b\)](#), and ICL’s performance can match that of large supervised models [García et al. \(2023\)](#). Subsequent work [Agrawal et al. \(2022\)](#) highlighted the importance of example quality in ICL. Instruction tuning [Wei et al. \(2022\)](#) has also been explored for translation. [Li et al. \(2023\)](#) fine-tuned XGLM-7B with translation-specific instructions to improve low-resource language performance, while [Chen et al. \(2021\)](#) enhanced instruction-following by introducing negative samples to improve translation faithfulness. Broadly, evaluation of LLM translation capabilities follows two lines: (1) LLM-specific factors, such as demonstration selection [García et al. \(2023\)](#); [Vilar et al. \(2022\)](#); [Zhang et al. \(2023\)](#) and prompt design [Jiao et al. \(2023\)](#) and (2) comprehensive benchmarking across multilingual [Hendy et al. \(2023\)](#); [Zhu et al. \(2023b\)](#), document-level

Method	Eng→Ben	Ben→Eng
Zero-shot	2.14	3.14
In context learning	1.96	2.96
Translation-CoT	1.39	2.16

Table 1: Averaged MQM score ($\downarrow$) for different methods on English–Bangla (Eng→Ben) and Bangla–English (Ben→Eng) translation. Lower is better.

Hendy et al. (2023); Karpinska and Iyyer (2023), low-resource García et al. (2023), robust Jiao et al. (2023), hallucination-resistant Guerreiro et al. (2023), domain-adaptive Wang et al. (2023a) settings and incorporated human evaluation Jiao et al. (2023) into instruction data, yielding translations preferred in interactive settings.

Chain-of-thought (CoT) prompting improves LLM performance on complex reasoning tasks, including arithmetic, commonsense, and symbolic reasoning, by eliciting intermediate reasoning steps before the final answer Wei et al. (2022); Wang et al. (2022). It is commonly implemented as zero-shot CoT, using triggers such as “Let’s think step by step” Yang et al. (2023), or as few-shot CoT, with demonstrations containing questions, reasoning chains, and answers Zhang et al. (2022). However, direct applications of CoT to machine translation remain limited, and naive adaptations can produce overly literal, word-by-word translations Peng et al. (2023). Relatedly, self-prompting decomposes inputs into sub-problems and extracts salient knowledge before solving the main task Li et al. (2022), with applications in in-context example generation Li et al. (2022), document generation Yu et al. (2022), summarization Wang et al. (2023b), and multimodal captioning Zhu et al. (2023a). He et al. (2023) introduced MAPS, which elicits keywords, topics, and relevant demonstrations before translation and uses quality estimation to select the final output. Briakou et al. (2024) proposed a multi-turn framework for long-form translation that decomposes the process into pre-translation research, drafting, refinement, and proofreading. In contrast, Translation-CoT performs lexical, grammatical, and topical analysis followed by refinement within a single structured prompt and evaluates this approach across multiple LLMs and language families.

6 Conclusion

Our study evaluates both the strengths and limitations of current LLMs in multilingual machine translation. Among the evaluated models, GPT-4o achieves the strongest overall performance, particularly for several high- and mid-resource languages, while substantial challenges remain in lower-resource settings. We introduce Translation-CoT, a prompting framework inspired by human translation workflows that combines lexical, grammatical, and topical analysis with a refinement phase targeting fluency, tone, idiomatic expression, and ambiguity resolution. Across the evaluated automatic metrics, Translation-CoT improves over Zero-Shot and In-Context Learning in many settings and achieves the strongest summarized BLEU performance for 8 of the 14 languages. However, the gains are not uniform, and Zero-Shot or In-Context Learning remains competitive for the other languages, depending on the model, translation direction, and evaluation metric.

Human evaluation on English↔Bangla further shows more favorable preference ratings, lower MQM penalty scores, and fewer hallucinations for Translation-CoT. Translation-CoT also surpasses NLLB in selected language directions, but it remains below Google Translate in all reported comparisons where Google Translate results are available. Overall, these findings suggest that structured, translation-specific reasoning can improve LLM translation quality in several evaluated settings, although it should not be interpreted as a universally superior prompting strategy or as a replacement for specialized machine translation systems.

7 Limitations

This study evaluates one language from each of 14 language families and focuses only on English-centric translation directions. Therefore, the results may not generalize to other languages, within-family

variation, or non-English-centric translation. The experiments are also limited to sentence-level translation and do not test document-level context. Human preference, MQM, and hallucination evaluations were conducted only for English $\leftrightarrow$ Bangla because of annotator availability. We also did not perform component-level ablations to isolate the effects of lexical retrieval, grammatical analysis, topic identification, and refinement. Future work should examine which stages contribute most to the observed gains.

Our automatic evaluation relies on BLEU, ChrF, and METEOR, which may not fully capture semantic adequacy or discourse-level quality. Future evaluation should include neural metrics such as COMET and MetricX. Because FLORES-101 has been publicly available since 2021, possible benchmark contamination in LLM pretraining data cannot be ruled out. Finally, we did not compare Translation-CoT with fine-tuned translation models or the most recently released LLMs. The results are also based on single deterministic runs and may still be affected by API-side nondeterminism or model updates.

Acknowledgments

The authors would like to thank the reviewers for their feedback on this project. Our team acknowledges support from the MassMutual Center of Excellence in Complex Systems and Data Science (Grant # FP2860), Alfred P. Sloan Foundation (Grant #G-2024-22498), and the National Science Foundation (Grant #2242829).

References

- Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., and Ghazvininejad, M. (2022). In-context Examples Selection for Machine Translation. In *Annual Meeting of the Association for Computational Linguistics*.
- Alves, D. M., Guerreiro, N. M., Alves, J., Pombal, J. P., Rei, R., de Souza, J. G. C., Colombo, P., and Martins, A. (2023). Steering Large Language Models for Machine Translation with Finetuning and In-Context Learning. In *Conference on Empirical Methods in Natural Language Processing*.
- Baker, M. (1992). In Other Words: A Coursebook on Translation.
- Banerjee, S. and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *IEEEvaluation@ACL*.
- Bassnett, S. and Lefèvre, A. (1990). Translation, History and Culture.
- Bowker, L. (2002). Computer-Aided Translation Technology: A Practical Introduction.
- Briakou, E., Luo, J., Cherry, C., and Freitag, M. (2024). Translating Step-by-Step: Decomposing the Translation Process for Improved Translation Quality of Long-Form Texts. In Haddow, B., Kocmi, T., Koehn, P., and Monz, C., editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T. J., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., teusz Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv*, abs/2005.14165.
- Chen, G., Ma, S., Chen, Y., Zhang, D., Pan, J.-Y., Wang, W., and Wei, F. (2021). Towards Making the Most of Cross-Lingual Transfer for Zero-Shot Neural Machine Translation. In *Annual Meeting of the Association for Computational Linguistics*.
- Dai, Y., Chan, C. F., Wong, Y. K., and Pun, T. H. (2025). Learning to Learn: When Chain-Of-Thought in Domain-Specific Machine Translation. *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.
- Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Li, L., and Sui, Z. (2022). A Survey on In-context Learning. In *Conference on Empirical Methods in Natural Language Processing*.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A. S., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Rozière, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., tian Cantón Ferrer, C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D.,

Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E. A., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I. M., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J.-Q., Alwala, K. V., Upasani, K., Plawiak, K., Li, K., neth Heafield, K.-., Stone, K. R., El-Arini, K., Iyer, K., Malik, K., ley Chiu, K., Bhalla, K., Rantala-Yearly, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M. H. M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., lay Bashlykov, N., Bogoychev, N., Chatterji, N. S., Duchenne, O., cCelebi, O., Alrassy, P., Zhang, P., Li, P., Vasić, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Girdhar, R., Patel, R., main Sauvestre, R., nie Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., hana Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S. C., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., ginie Do, V., Vogeti, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., ney Meers, W., Martinet, X., Wang, X., Tan, X. E., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A. K., Grattafiori, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Vaughan, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Franco, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, P.-Y. B., Loyd, B., de Paola, B., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Civin, D., Beaty, D., Kreymer, D., Li, S.-W., Wyatt, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Ozgenel, F., Caggioni, F., Guzm'an, F., Kanayet, F. J., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Thattai, G., Herman, G., Sizov, G. G., Zhang, G., Lakshminarayanan, G., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Molybog, I., Tufanov, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., KamHou, U., Saxena, K., Prasad, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Huang, K., Chawla, K., Lakhotia, K., Huang, K., Chen, L., Garg, L., Lavender, A., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Tsimpoukelli, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., ish Bansal, M., Santhanam, N., Parks, N., White, N., ata Bawa, N., Singhal, N., Egebo, N., Usunier, N., Laptev, N. P., Dong, N., Zhang, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., dro Rittner, P., Bontrager, P., Roux, P., Dollár, P., Zvyagina, P., Ratan-chandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Li, R., Hogan, R., Battey, R., Wang, R., Maheswari, R., Howes, R., Rinott, R., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Shankar, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S. K., Cho, S.-B., Virk, S., Subramanian, S., Choudhury,

- S., Goldman, S., Remez, T., Glaser, T., Best, T., Kohler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V. A., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wang, X., Wu, X., Wang, X., Xia, X., Wu, X., Gao, X., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Wang, Y., Hao, Y., Qian, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., and Zhao, Z. (2024). The Llama 3 Herd of Models. *ArXiv*, abs/2407.21783.
- Filippova, K. (2020). Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data. In *Findings*.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.
- Freitag, M., Foster, G. F., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. (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.
- García, X., Bansal, Y., Cherry, C., Foster, G. F., Krikun, M., Feng, F., Johnson, M., and Firat, O. (2023). The unreasonable effectiveness of few-shot learning for machine translation. *ArXiv*, abs/2302.01398.
- Ghazvininejad, M., Gonen, H., and Zettlemoyer, L. (2023). Dictionary-based Phrase-level Prompting of Large Language Models for Machine Translation. *ArXiv*, abs/2302.07856.
- Gile, D. (1995). Basic Concepts and Models for Interpreter and Translator Training.
- Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. (2021). The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Guerreiro, N. M., Alves, D. M., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., and Martins, A. (2023). Hallucinations in Large Multilingual Translation Models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.
- Hatim, B. A. and Munday, J. (2005). Translation: An Advanced Resource Book.
- He, Z., Liang, T., Jiao, W., Zhang, Z., Yang, Y., Wang, R., Tu, Z., Shi, S., and Wang, X. (2023). Exploring Human-Like Translation Strategy with Large Language Models. *Transactions of the Association for Computational Linguistics*, 12:229–246.
- Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., and Awadalla, H. H. (2023). How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *ArXiv*, abs/2302.09210.
- Iroda, K. and Risolat, I. (2025). Identifying Grammatical Issues in Translation: Methods and Prospects for Technical Education Students. *SHS Web of Conferences*.
- Jiao, W., Wang, W., Huang, J.-T., Wang, X., and Tu, Z. (2023). Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine.
- Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F. B., Wattenberg, M., Corrado, G. S., Hughes, M., and Dean, J. (2016). Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of Tricks for Efficient Text Classification. *ArXiv*, abs/1607.01759.
- Karpinska, M. and Iyyer, M. (2023). Large Language Models Effectively Leverage Document-level Context for Literary Translation, but Critical Errors Persist. In *Conference on Machine Translation*.
- Koehn, P. (2009). A process study of computer-aided translation. *Machine Translation*, 23:241–263.
- Li, J., Zhang, Z., and Zhao, H. (2022). Self-Prompting Large Language Models for Open-Domain QA. *ArXiv*, abs/2212.08635.
- Li, J., Zhou, H., Huang, S., Chen, S., and Chen, J. (2023). Eliciting the Translation Ability of Large Language Models via Multilingual Finetuning with Translation Instructions. *Transactions of the Association for Computational Linguistics*, 12:576–592.

- Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Lommel, A. (2013). Multidimensional Quality Metrics : A Flexible System for Assessing Translation Quality.
- Newmark, P. (1988). A textbook of translation.
- Nguyen, L. and Xu, Y. (2025). Reasoning for Translation: Comparative Analysis of Chain-of-Thought and Tree-of-Thought Prompting for LLM Translation. In *Annual Meeting of the Association for Computational Linguistics*.
- Oni, M. K. and Prama, T. T. (2025). Transformer-Based Low-Resource Language Translation: A Study on Standard Bengali to Sylheti. *ArXiv*, abs/2510.18898.
- OpenAI, J. A., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., ing Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., laine Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., abella Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, L., Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, H., Kiros, J. R., Knight, M., Kokotajlo, D., Kondraciuk, L., Kondrich, A., Konstantinidis, A., Koscic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C., Lim, R., Lin, M., Lin, S., teusz Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., drey Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D. P., Mu, T., Murati, M., Murk, O., M'ely, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Long, O., O'Keefe, C., Pachocki, J. W., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Pokorny, M., Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J. W., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M. D., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N. A., Thompson, M., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C. L., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., ing Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2023). GPT-4 Technical Report.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Annual Meeting of the Association for Computational Linguistics*.
- Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., and Tao, D. (2023). Towards Making the Most of ChatGPT for Machine Translation. In *Conference on Empirical Methods in Natural Language Processing*.
- Pfau, M. (2015). Memes of Translation: The spread of ideas in translation theory.
- Popovic, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *WMT@EMNLP*.

- Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. In *Conference on Machine Translation*.
- Prama, T. T. and Anwar, M. M. (2025). Sylheti to Standard Bangla Neural Machine Translation: A Deep Learning-Based Dialect Conversion Approach. In Bajaj, A., Madureira, A. M., and Abraham, A., editors, *Hybrid Intelligent Systems. HIS 2023. Lecture Notes in Networks and Systems*, volume 1224. Springer, Cham.
- Prama, T. T., Danforth, C. M., and Dodds, P. S. (2025a). BanglaMATH : A Bangla benchmark dataset for testing LLM mathematical reasoning at grades 6, 7, and 8. *ArXiv*, abs/2510.12836.
- Prama, T. T., Danforth, C. M., and Dodds, P. S. (2025b). LLMs for Low-Resource Dialect Translation Using Context-Aware Prompting: A Case Study on Sylheti. *ArXiv*, abs/2511.21761.
- Raffel, C., Shazeer, N. M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.
- Raunak, V., Menezes, A., Post, M., and Awadallah, H. H. (2023). Do GPTs Produce Less Literal Translations? *ArXiv*, abs/2305.16806.
- Ren, S., Wu, Y., Liu, S., Zhou, M., and Ma, S. (2020). A Retrieve-and-Rewrite Initialization Method for Unsupervised Machine Translation. In *Annual Meeting of the Association for Computational Linguistics*.
- Riviere, G. T. M., Pathak, S., Sessa, P. G., Hardin, C., Bhu-patiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ram'e, A., Ferret, J., Liu, P., Tafti, P. D., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stańczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.-B., Neyshabur, B., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C. A., Choquette-Choo, C. A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozińska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshv, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucińska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J. R., Gordon, J., Lipschultz, J., Newlan, J., Ji, J., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L. L., Usui, L., Sifre, L., Heuermann, L., cia Lago, L., McNealus, L., Soares, L. B., Kilpatrick, L., Dixon, L., Martins, L. L. B., Reid, M., Singh, M., Iverson, M., Gerner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., shad Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., chong Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R. A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S. M. M., Perrin, S., Arnold, S. M. R., bastian Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kociský, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R., and Andreev, A. (2024). Gemma 2: Improving Open Language Models at a Practical Size. *ArXiv*, abs/2408.00118.
- Robinson, N. R., Ogayo, P., Mortensen, D. R., and Neubig, G. (2023). ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages. *ArXiv*, abs/2309.07423.
- Schwenk, H., Wenzek, G., Edunov, S., Grave, E., and Joulin, A. (2019). CCMatrix: Mining Billions of High-Quality Parallel Sentences on the Web. In *Annual Meeting of the Association for Computational Linguistics*.
- Taia.io (2025). 9 Reasons The Language Translation Process Is So Hard.
- team, N., Costa-jussà, M. R., Cross, J., cCelebi, O., El-bayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S. L., Tran, C., Andrews, P. Y., Ayan, N. F., Bhosale, S., Edunov, S.,

- Fan, A., Gao, C., Goswami, V., Guzm'an, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation. *ArXiv*, abs/2207.04672.
- Terryn, A. R. and de Lhoneux, M. (2024). Exploratory Study on the Impact of English Bias of Generative Large Language Models in Dutch and French. In *HUMEVAL*.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. *ArXiv*, abs/2302.13971.
- Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., and Foster, G. F. (2022). Prompting PaLM for Translation: Assessing Strategies and Performance. *ArXiv*, abs/2211.09102.
- Wang, H., Wu, H., He, Z., Huang, L., and Church, K. W. (2021). Progress in Machine Translation. *Engineering*.
- Wang, S., Yang, C.-H. H., Wu, J., and Zhang, C. (2023a). Can Whisper Perform Speech-Based In-Context Learning? *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13421–13425.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E. H., and Zhou, D. (2022). Self-Consistency Improves Chain of Thought Reasoning in Language Models. *ArXiv*, abs/2203.11171.
- Wang, Y., Zhang, Z., and Wang, R. (2023b). Element-aware Summarization with Large Language Models: Expert-aligned Evaluation and Chain-of-Thought Method. In *Annual Meeting of the Association for Computational Linguistics*.
- Weaver, W. (1955). Rededication to Liberal Decency. *Science*, 121 3138:15A.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E. H., Xia, F., Le, Q., and Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv*, abs/2201.11903.
- Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- Yang, A., Zhang, B., Hui, B., Gao, B., Yu, B., Li, C., Liu, D., Tu, J., Zhou, J., Lin, J., Lu, K., Xue, M., Lin, R., Liu, T., Ren, X., and Zhang, Z. (2024). Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. *ArXiv*, abs/2409.12122.
- Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., and Chen, X. (2023). Large Language Models as Optimizers. *ArXiv*, abs/2309.03409.
- Yu, W., Iter, D., Wang, S., Xu, Y., Ju, M., Sanyal, S., Zhu, C., Zeng, M., and Jiang, M. (2022). Generate rather than Retrieve: Large Language Models are Strong Context Generators. *ArXiv*, abs/2209.10063.
- Zhang, J. and Zong, C. (2020). Neural machine translation: Challenges, progress and future. *Science China Technological Sciences*, 63:2028–2050.
- Zhang, Z., Zhang, A., Li, M., and Smola, A. J. (2022). Automatic Chain of Thought Prompting in Large Language Models. *ArXiv*, abs/2210.03493.
- Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., and Smola, A. J. (2023). Multimodal Chain-of-Thought Reasoning in Language Models. *Trans. Mach. Learn. Res.*, 2024.
- Zhu, D., Chen, J., Haydarov, K., Shen, X., Zhang, W., and Elhoseiny, M. (2023a). ChatGPT Asks, BLIP-2 Answers: Automatic Questioning Towards Enriched Visual Descriptions. *ArXiv*, abs/2303.06594.
- Zhu, W., Liu, H., Dong, Q., Xu, J., Kong, L., Chen, J., Li, L., and Huang, S. (2023b). Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. *ArXiv*, abs/2304.04675.

A Lists of Language

We evaluate 14 languages in this paper. Table 2 lists the name, ISO code, and language family of these languages.

B Step-by-Step Translation Process Using Translation-CoT

The Translation-CoT framework enables large language models (LLMs) to incorporate reasoning into the translation process. Figure 8 shows in the **preparation phase**, the model first extracts translation-related knowledge from the source sentence, including *lexicons*, *grammar*, and *topics*. Leveraging this knowledge, the LLM then generates the target translation in the **translation phase**. Finally, in the **post-processing phase**, *fluency and tone refinement* along with *idiomatic and ambiguity resolution* are applied to produce a translation that is more natural, coherent, and human-like.

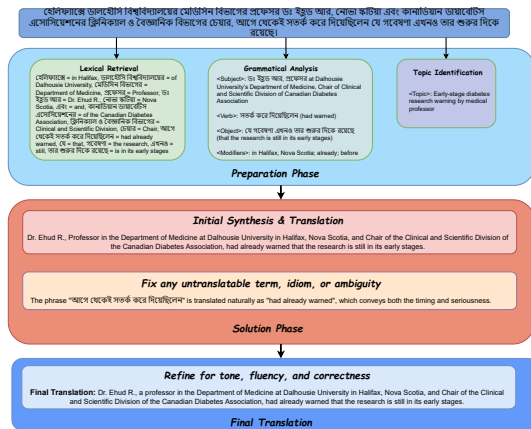


Figure 8: Example of translating a source sentence (source language: Bangla) into the target sentence (English) using the Translation-CoT reasoning steps (response from GPT-4o, accessed August 12, 2025).

C Prompt Strategies

We employ three prompting strategies for multi-lingual translation: Zero-Shot, In-Context Learning (ICL), and Translation with Chain-of-Thought (Translation-CoT). Below we describe each strategy with the exact prompt templates.

C.1 Zero-Shot Translation

The model is asked to translate directly from the source language to the target language without any examples:

Instruction:

You are an expert language translator. You are given sentences: {input_sentence} and translate the input from {source_language} to {target_language}.

Output: Return the final translated sentence in {target_language} only. Do not include any explanations or labels.

C.2 In-Context Learning (ICL)

The model is given a 5 source–target translation examples before translating the new sentence:

Instruction:

You are an expert language translator. You are given sentences: {input_sentence} and translate the input from {source_language} to {target_language}.

Here are a few examples of translation:

<Source> S1<Source><Target> T1 <Target>
<Source> S2<Source><Target> T2 <Target>
<Source> S3<Source><Target> T3 <Target>
<Source> S4<Source><Target> T4 <Target>
<Source> S5<Source><Target> T5 <Target>

Output: Return the final translated sentence in {target_language} only. Do not include any explanations or labels.

C.3 Translation with Chain-of-Thought (Translation-CoT)

The model follows a reasoning process that mimics human translation: Retrieving lexical meaning, analyzing grammar, identifying topics, translating, and refining:

Instruction:

You are an expert language translator. You are given sentences: {input_sentence}. Follow the steps one by one to translate the input from {source_language} to {target_language}.

Language	ISO 639-3	Family
Bangla	ben	Indo-European
Amharic	amh	Afro-Asiatic
Azerbaijani	azj	Turkic
Burmese	mya	Sino-Tibetan
Filipino (Tagalog)	tgl	Austronesian
Estonian	est	Uralic
Fula	ful	Atlantic-Congo
Georgian	kat	Kartvelian
Japanese	jpn	Japonic
English	eng	Indo-European
Kannada	kan	Dravidian
Korean	kor	Koreanic
Lao	lao	Kra-Dai
Luo	luo	Nilo-Saharan
Mongolian	mon	Mongolic

Table 2: Languages, ISO 639-3 codes, and their corresponding language families.

1. Retrieve the meaning of each word and phrase in the source sentence from a mental lexicon to the target language. If any term is ambiguous, note possible meanings.
2. Analyze the grammatical structure. Identify subject, verb, object, and modifiers.
3. Describe the topic of the sentence in a few words.
4. Translate using the identified knowledge.
5. Fix untranslatable terms or idioms by choosing the closest natural phrase in the target language.
6. Refine for tone, fluency, and correctness.

Output: Return the final translated sentence in {target_language} only. Do not include any explanations or labels.

D Translation performance of LLMs

For each of the 14 languages, we measured the BLEU, ChrF, and METEOR scores. Figures 9 and 10 present the performance of the LLMs, where scores are averaged across the three prompting techniques. For both English $\leftrightarrow$ Non-English (X) translation directions, GPT-4o consistently outperforms the other

LLMs.

E Comparison of LLMs’ performance across Languages and prompt strategies

For each of the 14 languages and each prompt strategy, we used three metrics to evaluate the performance of the LLMs. Figures 11 and 12 show that, among the prompt strategies, Translation-CoT outperforms the other two in both ChrF and METEOR scores. Additionally, Tables 3 and 4 provide detailed BLEU, ChrF, and METEOR results for each LLM and prompting strategy in the $X \rightarrow \text{Eng}$ and $\text{Eng} \rightarrow X$ translation directions, respectively.

F Comparison of Chain-of-Thought Prompting

Figure 13 compares Translation-CoT with ToT and LOT using ChrF and METEOR across both translation directions, showing that Translation-CoT generally achieves stronger lexical, structural, and semantic alignment across the evaluated languages. Table 5 further compares GPT-4o with Translation-CoT against Google Translate and NLLB using BLEU, showing that Translation-CoT surpasses NLLB for several languages but remains below Google Translate across the evaluated directions.

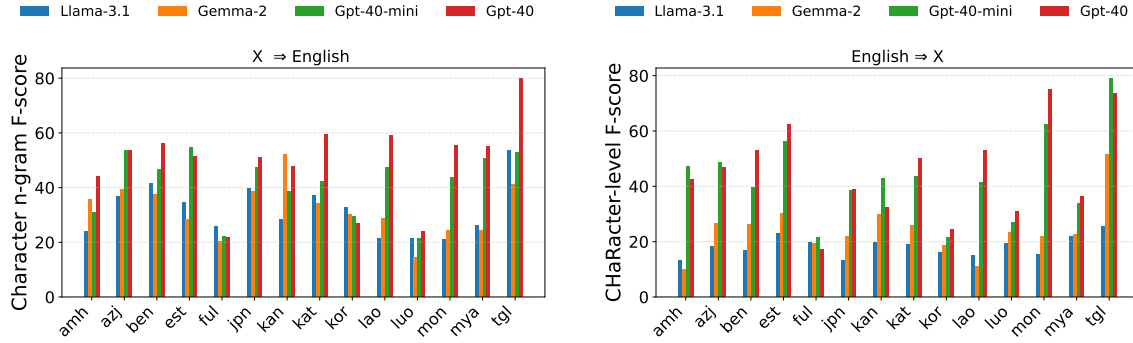


Figure 9: Translation performance (ChrF) of GPT-4o, GPT-4o-mini, LLaMA 3.1, and Gemma 2 across the evaluated languages. “X → Eng” and “Eng → X” denote translation into English and from English, respectively. ChrF scores are averaged over the three prompt strategies used in this experiment for each language.

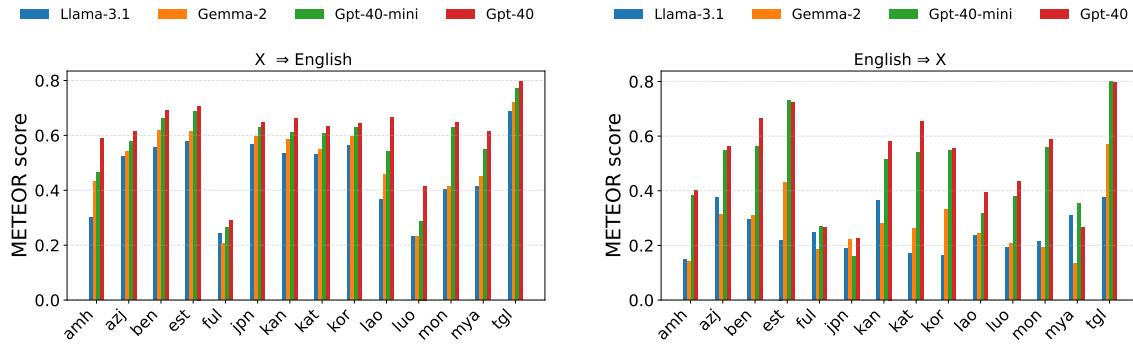


Figure 10: Translation performance (METEOR) of GPT-4o, GPT-4o-mini, LLaMA 3.1, and Gemma 2 across the evaluated languages. “X → Eng” and “Eng → X” denote translation into English and from English, respectively. METEOR scores are averaged over the three prompt strategies used in this experiment for each language.

G Human Evaluation Guidelines

This appendix presents the annotation instructions and rating criteria used in the human preference, Multidimensional Quality Metrics (MQM), and hallucination evaluations. Three graduate students from Bangladesh who were fluent in both Bangla and English independently completed the annotations. Participation was voluntary and uncompensated.

General Annotation Instructions For each sample, annotators were shown the source sentence together with translations generated using Zero-Shot, In-Context Learning, and Translation-CoT prompting. Annotators were instructed to evaluate each translation independently based on the meaning of

the source sentence and the quality of the corresponding translation. They were asked to assess whether the translation:

- accurately preserved the meaning of the source sentence;
- retained all important information from the source;
- avoided introducing information not present in the source;
- was grammatically correct in the target language;
- used natural and fluent expressions;

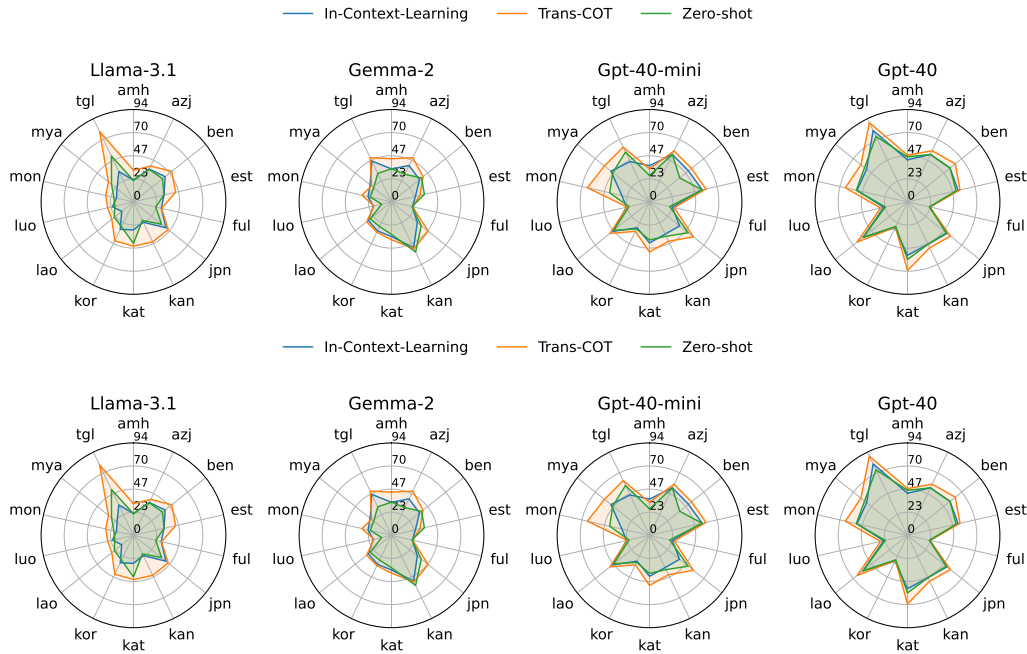


Figure 11: Multilingual translation performance (ChrF) of selected large language models (LLMs) for non-English (X) → English (top) and English → non-English (X) (bottom) translation across three prompt strategies: Zero-shot, In-context Learning, and Translation-CoT. In the majority of cases (8 out of 14 languages), the Translation-CoT prompt outperforms other prompt strategies in ChrF score for LLM performance.

- appropriately translated idiomatic, ambiguous, and culturally specific expressions; and
- maintained the tone and intent of the source sentence.

Annotators were instructed not to prefer a translation solely because it followed the source sentence literally. A natural translation that accurately preserved the original meaning was considered preferable to an awkward word-for-word translation.

Human Preference Rating Each translation was assigned one of three quality ratings: *Good*, *Fair*, or *Poor*. The rating criteria were defined as follows:

Good: The translation accurately preserves the meaning of the source sentence and is fluent, grammatically correct, complete, and natural in the target language. Minor stylistic differences that do not affect meaning or readability are acceptable.

Fair: The translation conveys the main meaning of the source sentence but contains one or more

noticeable issues, such as awkward phrasing, minor grammatical errors, slightly inaccurate word choices, or small omissions. These issues do not substantially alter the overall meaning.

Poor: The translation contains substantial errors that affect meaning or readability. These may include serious mistranslations, major omissions, untranslated content, hallucinated information, incorrect grammatical structures, or language that is difficult to understand.

When uncertain between two categories, annotators were instructed to assign the lower rating if the error affected the meaning of the sentence and the higher rating if the issue was primarily stylistic. Inter-annotator agreement for the *Good*, *Fair*, and *Poor* ratings was measured using Fleiss' κ . The resulting score of $\kappa = 0.78$ indicates substantial agreement among the three annotators.

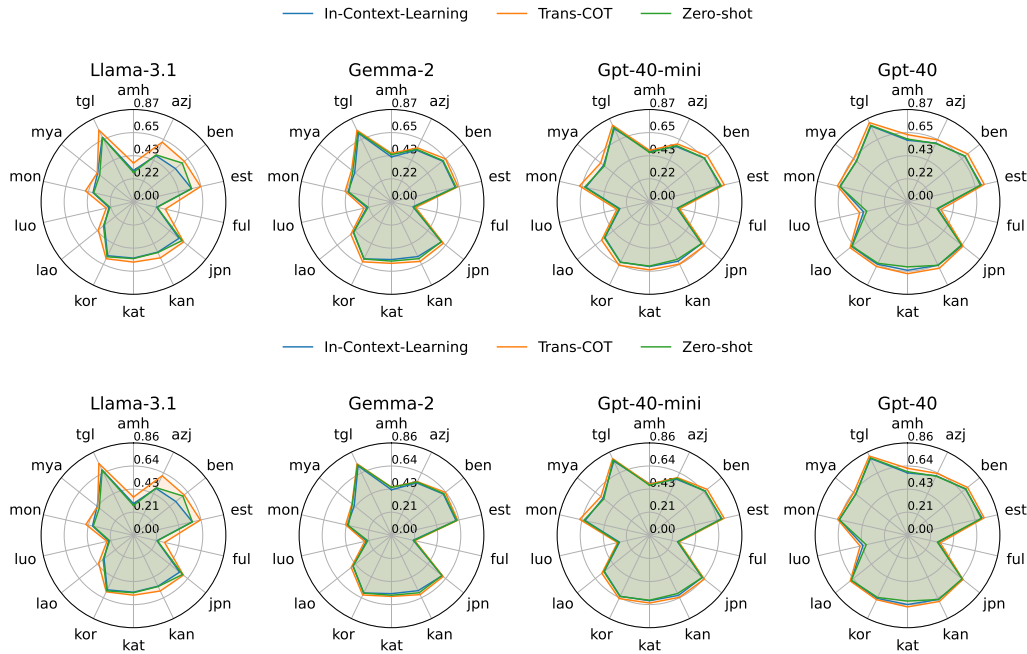


Figure 12: Multilingual translation performance (METEOR) of selected large language models (LLMs) for non-English(X) $\rightarrow$ English (top) and English $\rightarrow$ non-English (X) (bottom) translation across three prompt strategies: Zero-shot, In-context Learning, and Translation-CoT. In almost all cases (8 out of 14 languages), the Translation-CoT prompt outperforms other prompt strategies in METEOR score for LLM performance.

G.1 MQM Annotation Guidelines

The Multidimensional Quality Metrics (MQM) evaluation was conducted by the same three graduate students from Bangladesh who participated in the human preference evaluation.

For each sample, the annotators were shown the source sentence and its corresponding translated sentence. They were instructed to identify all translation errors, mark the affected text span, assign each error to one of the predefined categories, and determine its severity. A single translation could contain multiple errors from the same or different categories. The following four error categories were used:

Mistranslation: The translation conveys an incorrect meaning or inaccurately represents a word, phrase, relation, or proposition from the source sentence. This category includes incorrect lexical choices, incorrect interpretation of ambiguous expressions, and changes to the intended meaning.

Awkward style: The translation is understandable

but sounds unnatural, overly literal, repetitive, unclear, or inconsistent with common usage in the target language. This category includes inappropriate phrasing, unnatural word order, and stylistic constructions that reduce fluency.

Untranslated text: A word, phrase, or segment from the source language remains untranslated in the target-language output when a translation is required.

Omission: Information present in the source sentence is missing from the translation. An omission may involve a word, phrase, modifier, relation, or complete proposition that contributes to the source meaning.

Severity Levels Each identified error was assigned one of two severity levels:

Minor: The error is noticeable and may reduce fluency, naturalness, or precision, but it does not substantially alter the central meaning of the

source sentence. Minor errors received a penalty weight of 1.

Major: The error changes, obscures, or removes important information from the source sentence, produces a misleading interpretation, or substantially reduces the usability of the translation. Major errors received a penalty weight of 5.

Annotators were instructed to assign a *major* label when an error affected the meaning, accuracy, or completeness of the translation. Errors that primarily affected style or naturalness without substantially changing the meaning were generally labeled *minor*.

The MQM penalty score for each translation was calculated by summing the severity weights of all identified errors:

$$N_{\text{minor}} + 5N_{\text{major}} \quad (1)$$

where N_{minor} and N_{major} denote the numbers of

minor and major errors identified in the translation, respectively. Because a translation could contain multiple errors, the sentence-level MQM score did not have a fixed maximum value. The final score for each prompting strategy was calculated by averaging the sentence-level penalty scores across all evaluated samples. Lower MQM scores indicate fewer or less severe translation errors.

G.2 Hallucination Annotation Guidelines

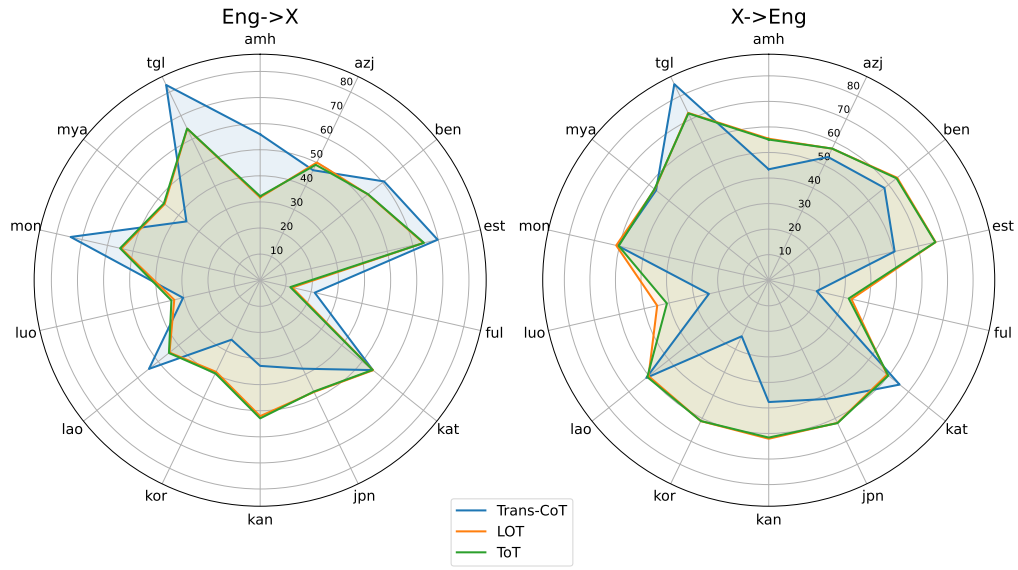
The same three bilingual annotators independently assigned each translation a binary label indicating the presence or absence of hallucination. A translation was labeled as hallucinatory if it introduced information, entities, events, attributes, or claims unsupported by or inconsistent with the source sentence. Paraphrasing, word-order differences, and grammatical or stylistic errors were not considered hallucinations unless they introduced unsupported meaning. Final labels were determined by majority vote.

Prompt	Language	LLaMA-3.1			Gemma-2			GPT-4o			GPT-4o-mini		
		BLEU	ChrF	METEOR	BLEU	ChrF	METEOR	BLEU	ChrF	METEOR	BLEU	ChrF	METEOR
Zero-shot	amh	7.2200	21.8700	0.2766	7.6600	33.6100	0.4441	25.9200	45.6600	0.5901	6.9600	26.8100	0.4639
	azj	12.4000	36.6300	0.4897	16.8100	31.0500	0.5480	24.1200	53.4000	0.6089	23.1200	53.4000	0.5863
	ben	12.3400	37.7100	0.5851	12.4500	39.0100	0.6206	33.2300	55.1000	0.6910	34.8900	39.1200	0.6591
	mya	8.3300	28.4700	0.4051	9.3500	23.0700	0.4645	16.5100	55.6700	0.6125	15.7000	47.6200	0.5415
	est	9.9800	31.7500	0.5601	8.3900	34.2200	0.6216	40.2100	50.5500	0.7016	41.6000	54.7400	0.6893
	tgl	16.7800	51.5800	0.6692	13.9100	32.0700	0.7260	53.1400	73.7000	0.7977	30.4600	56.0600	0.7775
	ful	6.1200	23.2200	0.2276	6.7400	21.3000	0.2127	6.4600	22.6400	0.2835	8.6700	21.0300	0.2694
	kat	30.8200	42.1100	0.5323	14.4000	32.8900	0.5557	21.4700	58.4200	0.6097	20.7500	38.5600	0.6020
	jpn	12.7400	36.3900	0.5827	14.7800	38.6000	0.5985	29.0400	50.2700	0.6471	30.2400	50.1600	0.6264
	kan	8.8500	20.8200	0.5272	12.0300	56.3800	0.5910	29.7600	47.6900	0.6606	14.8300	38.6300	0.5991
	kor	13.5900	27.9900	0.5716	13.8500	27.8300	0.5951	30.3800	27.5900	0.6396	26.3200	30.4400	0.6295
	lao	7.9100	24.8200	0.3604	10.8800	27.5700	0.4553	21.6400	57.6900	0.6665	16.5700	46.3300	0.5412
	luo	4.6700	19.4000	0.2262	6.0800	10.2000	0.2253	8.4800	23.0900	0.3927	4.9900	21.2700	0.2940
	mon	4.0700	20.4800	0.3829	5.1400	22.0300	0.4106	22.2800	53.2700	0.6494	21.1600	41.4400	0.6246
In-Context-Learning	amh	11.8900	22.0600	0.2942	8.1000	33.4300	0.4219	22.4900	42.6100	0.5789	11.8900	36.7500	0.4775
	azj	11.4500	36.9300	0.4830	17.2100	40.9700	0.5394	26.1200	53.4000	0.6129	24.1200	53.4000	0.5769
	ben	17.5300	41.0800	0.5026	12.5800	36.3000	0.6143	37.2300	55.1000	0.6874	37.8400	50.4600	0.6592
	mya	6.8900	21.3600	0.4310	4.0900	27.7000	0.4420	18.4600	53.0500	0.6142	18.0900	50.0100	0.5513
	est	9.9800	31.7500	0.5601	11.9400	24.8400	0.6139	40.7500	52.5900	0.7108	41.6000	54.7400	0.6839
	tgl	17.4000	34.2200	0.6742	17.8900	46.1200	0.7150	52.9400	80.3300	0.7907	42.5400	45.3900	0.7632
	ful	8.2200	28.9100	0.2236	3.2100	21.8000	0.2066	6.9900	23.4100	0.2912	7.5400	23.7000	0.2660
	kat	16.1900	28.4300	0.5274	16.1900	35.1400	0.5410	22.6300	54.0600	0.6418	22.2000	41.5600	0.6070
	jpn	11.1900	42.1100	0.5477	14.2200	33.4800	0.6015	29.4500	51.1600	0.6556	31.6500	38.9800	0.6262
	kan	8.3900	22.8900	0.5237	17.5700	51.5900	0.5696	29.7600	47.6900	0.6606	6.2900	36.3200	0.6183
	kor	9.7800	30.8800	0.5607	14.9500	32.6100	0.5933	33.2300	29.1600	0.6489	30.1600	29.1200	0.6276
	lao	5.8700	15.2900	0.3489	3.8800	30.0400	0.4579	21.6400	57.6900	0.6719	19.2200	48.0400	0.5361
	luo	5.4700	21.9400	0.2362	9.8400	18.8700	0.2373	8.6400	24.8500	0.4229	7.2600	23.9600	0.2854
	mon	4.9200	17.6200	0.3898	10.9000	24.3900	0.4190	22.6500	52.5200	0.6487	18.7700	28.7000	0.6131
Translation-CoT	amh	12.2300	28.1800	0.3348	13.8500	39.7600	0.4268	29.9300	43.4300	0.6014	12.6700	29.5900	0.4541
	azj	21.4100	36.3600	0.5930	27.3400	45.7400	0.5326	28.6800	53.5400	0.6169	25.1200	53.4000	0.5737
	ben	37.8800	45.6400	0.5811	22.5000	36.8900	0.6185	40.5100	57.9600	0.6920	37.8400	50.4600	0.6638
	mya	8.3300	28.4700	0.4051	11.3100	21.8200	0.4490	20.4400	56.4100	0.6175	18.5400	54.4300	0.5564
	est	29.4200	39.8200	0.6158	12.6100	25.5600	0.6097	42.2100	50.5500	0.7089	41.6000	54.7400	0.6902
	tgl	68.1500	75.2100	0.7192	34.8200	45.7200	0.7163	55.4600	85.1800	0.7959	44.2300	57.5800	0.7692
	ful	9.1200	25.2200	0.2760	5.6000	17.9600	0.2018	8.1000	19.2600	0.2909	9.4800	21.7500	0.2591
	kat	43.2400	40.8900	0.5351	21.9400	34.3500	0.5468	25.9600	65.5300	0.6439	23.5200	46.8500	0.6082
	jpn	17.1500	40.9900	0.5722	13.0400	43.5400	0.5929	30.0800	51.6200	0.6361	31.2400	52.8200	0.6308
	kan	12.8300	40.9900	0.5529	19.5500	48.4600	0.5902	29.7600	47.6900	0.6606	18.4900	40.3200	0.6183
	kor	19.4500	39.7400	0.5618	16.4400	30.1100	0.5945	33.8300	24.3600	0.6426	32.1600	29.1200	0.6306
	lao	8.3600	24.1100	0.3932	10.3800	28.2200	0.4580	25.7600	61.3000	0.6596	21.6000	47.3200	0.5452
	luo	7.3900	22.7700	0.2320	7.8600	14.2400	0.2347	10.2700	24.0200	0.4309	12.7900	19.2400	0.2844
	mon	11.0200	24.8900	0.4327	12.9000	26.7600	0.4163	24.7300	60.6300	0.6450	27.7300	60.6300	0.6441

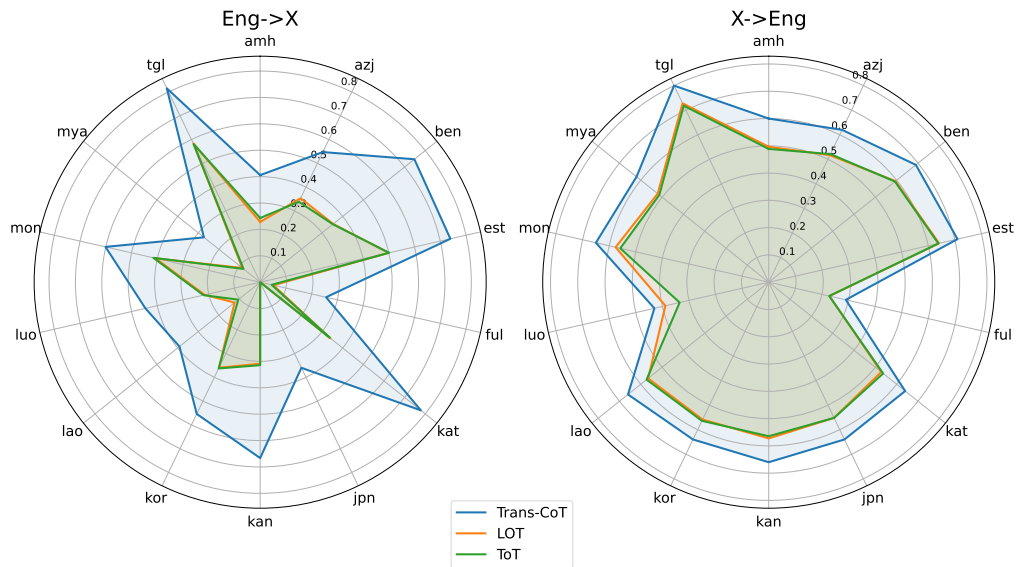
Table 3: Translation performance (BLEU, ChrF, METEOR) of GPT-4o, GPT-4o-mini, LLaMA-3.1, and Gemma-2 across 14 evaluated languages for “X→Eng” translation. Scores are the average of each test set for each language, measured using BLEU, ChrF, and METEOR metrics. Blue shading indicates that Translation-CoT outperformed both In-Context Learning and Zero-Shot for that metric. Gray shading marks cases where Zero-Shot scored highest among the three, while orange shading highlights In-Context Learning as the best performer. Value with no shading indicate a tie across methods for that metric. Higher scores reflect better translation performance.

Prompt	Language	LLaMA-3.1			Gemma-2			GPT-4o			GPT-4o-mini		
		BLEU	ChrF	METEOR	BLEU	ChrF	METEOR	BLEU	ChrF	METEOR	BLEU	ChrF	METEOR
Zero-shot	amh	5.2386	14.7144	0.1596	7.4586	11.6892	0.1371	12.6749	33.2422	0.3877	18.4065	55.8916	0.4057
	azj	7.2769	18.4966	0.4623	8.4538	26.2514	0.3463	23.4856	46.6112	0.5697	22.0856	46.5912	0.5697
	ben	12.1515	14.3295	0.2184	12.2260	27.2624	0.3416	24.9474	48.8522	0.6471	18.5474	38.8322	0.5471
	mya	3.2730	19.5861	0.1747	4.0000	22.5562	0.1299	14.4384	36.8715	0.2647	13.0384	36.8515	0.2647
	est	5.1882	20.5951	0.1694	11.1239	29.6062	0.5400	34.9824	60.1519	0.7457	33.5824	60.1319	0.7457
	tgl	4.6297	18.6886	0.1715	23.2599	47.9698	0.6466	37.8293	76.8564	0.8010	34.4293	76.8364	0.8010
	ful	3.7317	16.2046	0.1735	8.4569	18.3448	0.1969	5.7562	9.4835	0.2588	12.0931	21.4204	0.2561
	kat	3.9089	13.6787	0.1427	13.8375	22.4866	0.2936	13.2846	49.4192	0.5871	11.8846	49.3992	0.5871
	jpn	2.1363	17.0335	0.2455	4.3600	32.5460	0.2100	5.2000	34.5236	0.1600	3.8000	34.5036	0.1600
	kan	3.7970	16.0876	0.3771	10.5117	40.7354	0.3204	17.2786	32.2528	0.5427	19.8786	32.2328	0.5427
	kor	4.3700	20.2637	0.1548	14.3436	19.0009	0.3801	14.4185	24.0620	0.5595	13.0185	24.0420	0.5595
	lao	2.7569	15.0157	0.1772	4.0000	4.0500	0.2633	11.4747	47.9242	0.3809	10.0747	47.9042	0.3809
	luo	1.6636	19.8440	0.1716	8.0272	25.0294	0.2151	11.1080	33.2069	0.4527	9.7080	33.1869	0.4527
	mon	1.9430	1.1141	0.0277	11.4904	25.8604	0.2057	9.5003	73.2403	0.6089	9.1003	73.2203	0.6089
In-Context-Learning	amh	6.5016	6.3898	0.0262	4.3133	8.4093	0.1351	11.6685	37.7719	0.4060	7.9130	25.7767	0.3070
	azj	7.3246	17.3057	0.4608	5.6112	22.8219	0.2463	23.3536	46.7782	0.5714	23.9434	49.9991	0.5380
	ben	14.1150	15.5449	0.4332	24.8355	15.7948	0.2536	26.4961	48.8522	0.6031	19.3507	39.2019	0.5196
	mya	5.1412	22.9780	0.3914	4.5540	22.6000	0.1339	9.9892	35.5428	0.2594	12.7138	35.3702	0.2713
	est	6.1775	21.8241	0.1968	4.9796	25.0137	0.2395	32.9824	56.6563	0.6845	34.6975	48.3024	0.6975
	tgl	6.2237	19.8162	0.4822	7.5688	33.5820	0.4325	38.9581	60.6289	0.7755	35.4462	82.7910	0.7931
	ful	5.5583	20.2424	0.4980	4.4855	17.6939	0.1825	7.7624	20.8398	0.2831	11.3358	22.2027	0.2722
	kat	4.3699	17.7012	0.1720	6.4165	23.8660	0.2138	13.7484	45.2384	0.5948	11.8393	39.8806	0.5170
	jpn	3.5482	10.3976	0.1443	4.3236	8.8106	0.1024	5.2000	44.5212	0.1600	3.8000	41.4120	0.1600
	kan	4.6535	20.5404	0.5200	6.2581	19.3653	0.2069	20.0013	32.0248	0.5376	18.4645	49.4230	0.4972
	kor	1.8194	7.6992	0.1394	5.0710	10.5191	0.2414	14.4185	24.0620	0.5523	20.7436	24.5180	0.5494
	lao	3.1600	15.0250	0.3880	4.8555	8.6581	0.2219	12.4747	56.8985	0.4123	10.3673	38.3138	0.2958
	luo	1.9823	19.0372	0.1003	4.3445	25.0100	0.1881	11.9167	28.6885	0.4051	12.1552	24.6382	0.3377
	mon	5.3286	22.2396	0.0985	8.3545	15.7046	0.1594	11.3654	77.0625	0.5571	8.9411	68.6978	0.5250
Translation-CoT	amh	7.9083	18.6945	0.2589	9.6698	9.3980	0.1494	19.8065	55.9116	0.4057	19.8650	59.5856	0.4345
	azj	15.6551	18.6407	0.2065	17.5086	30.8306	0.3445	24.3536	46.7782	0.5473	25.4356	49.7168	0.5321
	ben	26.3523	21.0742	0.2341	27.5064	36.0597	0.3315	29.0561	60.7262	0.7473	21.6226	40.2186	0.6221
	mya	5.2000	23.3996	0.3653	8.7677	23.0640	0.1329	14.0892	36.1350	0.2726	17.3235	29.5983	0.5277
	est	7.1270	25.9493	0.2932	9.8641	35.9448	0.5124	36.4815	69.7704	0.7393	33.5824	60.1319	0.7457
	tgl	8.3672	37.6469	0.4733	32.4333	72.7265	0.6279	40.5564	83.0445	0.8156	34.4293	76.8364	0.8010
	ful	9.8166	22.1473	0.0687	10.0196	21.7357	0.1749	8.4931	21.4404	0.2561	14.8168	20.5132	0.2789
	kat	7.1418	25.3149	0.1976	10.4685	30.6767	0.2788	17.4843	55.2384	0.7789	12.6475	41.2311	0.5125
	jpn	6.5600	12.0292	0.1759	6.2100	24.6824	0.3500	8.2000	37.6255	0.3600	3.8000	39.1712	0.1600
	kan	5.0313	22.3117	0.1958	12.6979	29.1496	0.3104	22.9190	32.8322	0.6664	21.1620	46.8444	0.4996
	kor	14.9911	20.1076	0.1979	13.8375	25.9696	0.3763	15.8258	25.2951	0.5540	24.3567	15.5864	0.5342
	lao	5.6920	15.0250	0.1388	4.0000	20.7713	0.2483	12.4747	54.4871	0.3910	11.6098	37.4489	0.2778
	luo	5.9823	19.0372	0.3010	7.9297	20.3195	0.2131	11.7185	30.2935	0.4459	12.1298	23.4696	0.3444
	mon	8.3286	22.2396	0.5130	12.4330	24.4322	0.2077	11.9335	74.4895	0.6012	11.8271	44.8540	0.5365

Table 4: Translation performance (BLEU, ChrF, METEOR) of GPT-4o, GPT-4o-mini, LLaMA-3.1, and Gemma-2 across 14 evaluated languages for “Eng→X” translation. Scores are the average of each test set for each language, measured using BLEU, ChrF, and METEOR metrics. Blue shading indicates that Translation-CoT outperformed both In-Context Learning and Zero-Shot for that metric. Gray shading marks cases where Zero-Shot scored highest among the three, while orange shading highlights In-Context Learning as the best performer. Value with no shading indicate a tie across methods for that metric. Higher scores reflect better translation performance.



(a) ChrF



(b) METEOR

Figure 13: Comparison of CoT-based prompting strategies (Translation-CoT, ToT, and LOT) for multilingual translation performance across 14 languages from diverse language families using GPT-4o, evaluated using ChrF (top) and METEOR (bottom) metrics for both translation directions (left: $\text{Eng} \rightarrow X$, right: $X \rightarrow \text{Eng}$).

Direction	Language	Google	NLLB	Translation-CoT	CoT vs NLLB	CoT vs Google
X→Eng	Amharic	38.99	32.98	29.93	—	—
	Bangla	43.37	36.97	40.51	+9.58%	—
	Azerbaijani	31.77	28.45	28.68	+0.81%	—
	Burmese	34.06	30.90	20.44	—	—
	Filipino	57.79	51.78	55.46	+7.11%	—
	Estonian	44.49	36.78	42.21	+14.76%	—
	Fula	—	—	8.10	—	—
	Georgian	33.96	29.94	25.96	—	—
	Japanese	36.68	31.67	30.08	—	—
	Kannada	39.33	36.89	29.76	—	—
	Korean	35.85	30.55	33.83	+10.74%	—
	Lao	43.33	37.49	25.76	—	—
	Luo	—	27.48	10.27	—	—
	Mongolian	38.39	29.47	24.73	—	—
Eng→X	Amharic	32.55	29.12	19.81	—	—
	Bangla	37.66	34.31	29.06	—	—
	Azerbaijani	25.92	21.23	24.35	+14.71%	—
	Burmese	27.10	18.66	14.09	—	—
	Filipino	41.83	37.46	40.56	+8.28%	—
	Estonian	41.82	32.73	36.48	+11.46%	—
	Fula	—	—	8.49	—	—
	Georgian	36.06	30.67	17.48	—	—
	Japanese	42.90	23.98	8.20	—	—
	Kannada	43.46	37.47	22.92	—	—
	Korean	31.96	28.08	15.83	—	—
	Lao	30.20	32.10	12.47	—	—
	Luo	—	—	11.72	—	—
	Mongolian	31.72	21.34	11.93	—	—

Table 5: BLEU comparison across Google Translate, NLLB, and GPT-4o Translation-CoT for both translation directions. Percentage gains are shown only where Translation-CoT outperforms the corresponding baseline.

Toward Equitable Machine Translation for Atypical Speech: An LLM Post-Correction Approach

Grace Pasion
Ammon Shurtz
Stephen D. Richardson

Department of Computer Science, Brigham Young University

gpasion0@byu.edu
acshurtz@byu.edu
srichardson@byu.edu

Abstract

Individuals with speech disabilities rely on everyday technologies powered by Automatic Speech Recognition (ASR) systems, yet these systems consistently fail them—producing significantly higher error rates that undermine the usefulness of voice assistants, hands-free devices, and machine translation pipelines. We conduct a multi-stage evaluation of speech impairment effects in cascaded speech-to-text translation, examining two distinct conditions: real dysarthric speech and simulated rhotacism. For dysarthria, we quantify ASR error rates. For rhotacism, we quantify error rates from a minimal-pair text substitution. We then analyze how impairment-induced errors propagate through downstream machine translation across three language directions (English to Spanish, Ukrainian, Khmer), and propose a training-free LLM-based post-correction methodology as an accessible intervention. We find that larger LLMs (70B parameters or more) consistently improve downstream translation quality even when surface-level corrections made by those LLMs remain modest, while smaller models lack the capacity to do so reliably. These results reveal a promising but scale-dependent path toward more equitable speech technology for users with atypical speech.

1 Introduction

Voice-user interfaces (VUIs) have become deeply embedded in everyday life, powering applications ranging from virtual assistants to accessibility tools. These technologies rely on sophisticated Automatic Speech Recognition (ASR) systems to interpret user intent. However, for a significant portion of the global population, these systems fail to perform reliably due to speech disorders and disabilities. According to a 2012 study, approximately 10% of U.S. adults have a speech, language, and/or voice disability (Morris et al., 2016), and more than 80 million people worldwide stutter.¹ Despite this substantial user base, those with speech disorders continue to encounter significantly higher error rates when using ASR-interfaced systems (Jefferson, 2019; Tobin et al., 2024), highlighting a critical gap between the populations these technologies serve and those they

are designed to accommodate.

ASR is an essential component for a range of downstream NLP tasks that rely on spoken input. One prominent example is Speech-to-Speech Machine Translation (S2ST), in which a cascaded pipeline integrates ASR, Machine Translation (MT), and Text-to-Speech (TTS) components to produce a fully end-to-end spoken translation system. In such pipelines, the performance of each component is inherently dependent on the quality of its predecessor, meaning that ASR errors propagate downstream and compound throughout the system (Matusov et al., 2005; Min et al., 2025). This cascade effect makes atypical speech particularly problematic: recognition errors introduced at the ASR stage can fundamentally distort the meaning of subsequent translations, disproportionately affecting speakers with speech disorders.

¹<https://www.stutteringhelp.org/faq>

While cascaded S2ST systems have seen considerable advancement in recent years, the robustness of such pipelines to atypical speech remains poorly understood. Despite substantial progress in ASR robustness, speech impediment modeling, and LLM-based error correction, existing work largely studies these components in isolation. Additionally, prior research on speech-impaired ASR output primarily focuses on evaluating recognition accuracy without examining how impairment-induced errors propagate to downstream tasks, like S2ST. Similarly, work on MT robustness to ASR noise typically assumes generic or synthetic noise patterns rather than systematic articulatory errors caused by speech impairments.

Large Language Models (LLMs) have recently emerged as a promising post-processing solution for ASR error correction (Pu et al., 2023; Tang et al., 2025; Iudin et al.), offering a potential remedy that does not require retraining or adapting the underlying recognition model. These studies rarely consider speech impairments explicitly, compare correction at different stages of the pipeline, or evaluate downstream translation quality. Whether input corrections effectively mitigate the compounding effects of speech-impairment noise in a cascaded translation pipeline remains an open question.

Thus, in this work, we aim to address these gaps through a systematic, multi-stage evaluation of speech-impairment effects in cascaded speech-to-text translation. We analyze both real dysarthric speech and a simulated rhotacism condition, comparing three open-source LLMs of varying sizes in their ability to correct impairment-induced ASR errors. A simulated rhotacism dataset is employed due to the lack of high-quality real datasets for this specific speech impairment. We then quantify how these corrections affect downstream translation quality from English into three target languages representing high, medium, and low-resource settings.

Overall, all three LLMs struggle to substantially correct dysarthria-induced ASR errors, achieving only up to a 1.4% relative improvement in Word Error Rate (WER) (Marzal and Vidal, 2002) compared to the original ASR output. In contrast, simulated rhotacism errors yield a modestly higher average improvement of up to 2.6% WER. Despite these limited gains at the recognition level, we observe a more nuanced pattern downstream. For both

dysarthric speech and simulated rhotacism, the two largest open-source LLMs evaluated — LLAMA3 (70B) and Qwen3 (235B) — consistently improve translation performance across all three language directions. In contrast, the smallest model (OLMo-3 7B) does not demonstrate sufficient capacity to produce consistent downstream gains. These findings suggest that modest surface-level ASR improvements can nonetheless yield measurable translation benefits, but that model scale plays a critical role in effective correction.

We release the code and datasets used here: <https://github.com/byu-matrix-lab/project-atypical-speech-mt>.

2 Related Works

2.1 Robustness of ASR

Automatic speech recognition (ASR) systems are sensitive to intrinsic speaker variations, such as pitch and speaking rate (Meyer et al., 2011), and to extrinsic acoustic noise, such as the background noise generated by public crowds (Moser et al., 2025).

Early work on ASR robustness focused on auditory processing-based features (Maganti and Matasoni, 2014), spectrum enhancement with sparse coding (He et al., 2015), and LSTM-based speech enhancement with feature fusion (Weninger et al., 2015). More current approaches leverage unsupervised self-training for noise and accent robustness (Hu et al., 2024), context-aware noise representation learning (Lee et al., 2024), and two-stage deep neural network spectral mapping (Wang et al., 2020). Other approaches focus on error correction via machine translation (Mani et al., 2020), error corrections via transformer-based models (Lin et al., 2025), and training-free intelligibility-guided fusion (Li et al., 2026) to make ASR more robust.

While these techniques are generally effective, they remain less effective when systematic errors arise due to speech impediments, necessitating further downstream correction and robustness strategies.

2.2 Robustness of MT to ASR Errors

ASR errors propagate directly to MT systems in traditional cascaded speech-to-text translation. Consequently, these errors often degrade translation quality.

Early work explored adapting an NMT model on ASR transcripts to improve model robustness towards these errors (Di Gangi et al., 2019). More

recent research proposed a unified framework, Whisper-UT, that improves robustness in multi-modal MT by simulating ASR errors (Xiao et al., 2025). Other work has focused on mitigating ASR errors in speech-to-text translation through cascaded systems that incorporate multiple ASR candidates and self-supervised speech features (Min et al., 2025). Additionally, some approaches for end-to-end models use regularized cross-attention to transfer alignment knowledge from ASR and MT tasks (Zhao et al., 2024). Finally, low-resource approaches exist that leverage synthetic data and model regularization to improve both cascaded and end-to-end ST performance (Li et al., 2025).

2.3 ASR Errors Induced by Speech Impediments

ASR performance degrades substantially when applied to speech affected by disorders such as dysarthria, apraxia, or articulation impairments. For example, commercial ASR platforms are limited in their ability to decipher dysarthric speech, for which a 80%-90% error rate has been reported (De Russis and Corno, 2019). Other research emphasizes the gap between ASR errors from users with speech impediments and those without one (Jefferson, 2019; Tobin et al., 2024). Clearly, there is work to be done to improve ASR models for users with disabilities.

2.4 Speech Impediment Correctors

Several methods have been developed to improve ASR performance for individuals with speech impairments. Techniques include EEG-assisted deep learning (Krishna et al., 2021), neural network-based speech assistance (Mounir et al., 2017), cross-lingual self-supervised representations (Hernandez et al., 2022), and two-stage data augmentation (Bhat and Strik, 2025). Further, other advances leverage transformer architectures (Shahamiri et al., 2023), empirical mode decomposition with CNNs (Sidi Yakoub et al., 2020), acoustic and lexical model adaptation (Mengistu and Rudzicz, 2011), hybrid feature extraction with evolutionary optimization (Rughani and Shivakrishna, 2015), and speaker-dependent parameter tuning (Marini et al., 2021) to improve recognition accuracy and robustness. Finally, methods such as fine-tuning the Wav2Vec 2.0 transformer model (Sapkota et al., 2026) and Mel spectrogram-based transfer learning for dysarthric ASR (Dhanalakshmi et al., 2025) further improve accuracy.

2.5 Large Language Models as ASR Correctors

Large Language Models (LLMs) have recently been applied as post-processing tools to improve ASR output. For general ASR errors, studies have explored multi-stage reasoning approaches (Pu et al., 2023), full-text error correction (Tang et al., 2025), post-correction fine-tuning (Iudin et al.), task-activating prompting (Yang et al., 2023), and zero-shot correction (Ma et al., 2025).

In addition to general ASR errors, LLMs have been used to correct errors in speech-impairment contexts, including dysarthric and stuttering speech. For example, La Quatra et al. (2025); Aboeitta et al. (2025); Bhat and Strik (2025); Lou et al. (2025) used LLM-based post-correction to improve speech-impaired ASR, while Hui et al. (2025) applied LLM-augmented methods in AAC software for dysarthric speakers. LLMs have also been applied to related tasks such as stuttering detection (Huang et al., 2025). Others have used either user-adapted, fine-tuned, or instruction-tuned LLMs to correct disordered speech (Hsu and Chang, 2025; Hernandez et al., 2025b,a). All in all, through previous research, LLMs have been shown to be robust at speech corrections for both ASR errors and ASR speech impediment errors.

3 Datasets

We employ a combination of clinical, simulated, and multilingual datasets to study speech impairments and induced rhotic errors.

3.1 Dysarthria

For our experiments, we use the Torgo dataset which contains audio recordings from individuals with speech impairments due to cerebral palsy or amyotrophic lateral sclerosis (Rudzicz et al., 2012). The dataset includes those with dysarthria and those without a speech impairment. The audio from those in the control group was filtered out. Additionally, the dataset was obtained by using a head mounted and directional microphone from the same session. We chose to use the head mounted data to better approximate real world conditions. Moreover, because Torgo consists of individual words and phrases, we extracted only the complete sentences for analysis.

Sentence transcription was performed using a large Whisper model, which is a state-of-the-art open-source ASR system widely used in research (Radford et al., 2023). After aligning the ASR output with the

original prompts, we were left with 691 lines of text.

3.2 Rhotacism

Current rhotacism datasets have their limitations, as many do not include the original prompts or full sentences. Consequently, we instead chose to simulate rhotic errors via text-based minimal-pair substitution. While designed to model potential ASR performance, this synthetic data does not represent genuine ASR transcripts.

3.2.1 Language Pairs

For our simulation, we use parallel sentences pulled from the Tatoeba¹ corpus as described by Tiedemann (2020). Tatoeba was selected because its large parallel corpora allowed us to filter for sentences containing sufficient rhotic phonemes while preserving aligned reference translations. In contrast, smaller benchmarks would have provided fewer usable examples after rhotacism-specific filtering. Additionally, Tatoeba is widely used in machine translation research, making it a suitable choice for our study. We pulled a random subset from three different datasets which include the English-to-Spanish, English-to-Ukrainian, and English-to-Khmer language pairs. Spanish, Ukrainian, and Khmer were chosen to compare the possible differences between a high-resource, a medium-resource, and a low-resource language.

To ensure the datasets were suitable for modeling induced rhotic errors, we filtered the data to retain only lines containing five “r”s in the English-Spanish and English-Ukrainian files, and two “r”s per line for English-Khmer. The resulting datasets included 1000 lines of English-to-Spanish, 999 lines of English-to-Ukrainian, and 386 lines of English-to-Khmer lines.

3.2.2 Simulation Techniques

Rhotacism is defined as a speech impediment that affects the production of the phoneme /r/, frequently leading to substitutions that resemble /l/ or /w/ (Flipsen Jr, 2021). Thus, in order to simulate rhotacism we utilize minimal pairs which are words that differ by one phoneme. For example, right (R AY1 T) and white (W AY1 T) are considered minimal pairs. To craft a minimal pair dictionary, we leverage CMUDict which is an open-source pronunciation dictionary (Carnegie Mellon University Speech Group, 2014).

To generate the simulated-rhotacism dataset, we first processed the original English sentences and identified minimal pairs within each sentence. Whenever a matching word was found, it was replaced with its corresponding minimal pair. To replicate real-world noise, there was a 1% probability that a word would remain unchanged. This simulation resulted in three aligned files of the original English-Target language pair and the aligned English rhotacism simulated file.

4 Methodology

We now define the technical architecture, algorithms, and data flow used to evaluate how speech-impairment-induced ASR errors affect downstream machine translation, and whether input-level correction mitigates these effects.

4.1 System Architecture Overview

Our system follows a cascaded speech-to-text translation pipeline consisting of four primary components:

1. Speech Recognition or Text Simulation
2. Optional Input Correction
3. Machine Translation
4. Evaluation

For dysarthria, the pipeline begins with impaired speech audio, while for rhotacism, it begins with clean English text that is algorithmically modified to simulate articulatory errors. In both cases, the output is English text that contains systematic recognition or articulation errors. Additionally, this text may optionally be corrected before being passed into a machine translation model. The translated outputs are then evaluated using automatic metrics.

4.2 Input-Level LLM Correction

To mitigate impairment-induced errors prior to translation, we apply input-level correction using instruction-tuned large language models. Each ASR output or simulated sentence is processed independently using a fixed prompt that instructs the model to correct errors caused by the specific speech impairment. The prompt used is included in Appendix A.

We evaluate three open-source LLMs:

- LLaMA-3.3-70B-Instruct
- Qwen3-235B-A22B

- OLMo-3-7B-Instruct

These models were selected to represent current state-of-the-art open-source systems across different parameter scales and training strategies (Grattafiori et al., 2024; Yang et al., 2025; Olmo et al., 2025).

LLMs are known to produce variable outputs or be sensitive to prompt formatting even under constrained prompting, with minor changes in prompt wording or format leading to substantially different responses across multiple models (Chatterjee et al., 2024). Thus, it is necessary that we apply deterministic post-processing rules. If the output has tags or formatting difficulty such as "*The corrected sentence is: [THE SENTENCE]*," it was cleaned to not include the extra format. Additionally, if the LLM output has no meaningful sentence to extract or there was a failure for any other reason such as model refusal for inappropriate sentences, the original sentence is returned. Finally, if the outputted sentence is a legible, extractable sentence it is simply returned. These rules applied to all correction files.

As a result from the input correction process, our erroneous English sentences were transformed to corrected English sentences.

4.3 Machine Translation

To evaluate downstream effects on translation, all English source variants are translated using the same MT system: facebook/nllb-200-distilled-600M (Costa-Jussà et al., 2022). Decoding uses the Hugging Face default generation settings (greedy decoding; no beam size or length penalty was explicitly specified), and these settings are held constant across all conditions. This model was selected due to its strong multilingual performance, coverage of low-resource languages, and open-source availability.

We translate all source conditions into Spanish, Ukrainian, and Khmer, representing high, medium, and low resource target languages. Translation is performed using identical decoding settings across all conditions to ensure comparability.

5 Experiments

5.1 Experimental Setup

Baselines and Comparisons For our experiments, we compare the following conditions:

- Clean Prompt Translation (Upper Bound): The original English Prompt translated directly
- Raw Impaired Translation (Baseline): The uncorrected ASR output or simulated rhotacism text translated
- Input-Corrected Translation: The LLM-corrected English input translated

This design isolates the effect of speech impairment and evaluates whether correction prior to MT recovers translation quality.

Data Splits and Processing All datasets are evaluated in a sentence-level, one-pass setting without train/dev/test splits, as no model parameters are trained. All processing is deterministic given fixed prompts and decoding parameters.

5.2 Evaluation Methodology

5.2.1 Speech Impediment Effect

To measure the severity of impairment-induced text corruption, we compare raw impaired text against the original English prompts using Word Error Rate (WER) which is a normalized edit distance computed from insertions, deletions, and substitutions between two sequences (Marzal and Vidal, 2002). WER directly measures recognition errors common in impaired speech. This metric allows us to quantify impairment severity and assess the effectiveness of input-level correction independently of translation.

5.2.2 Machine Translation Effect

For rhotacism, we compute BLEU scores using the sacrebleu library (Post, 2018). System outputs are compared against aligned reference translations within the same language. For our evaluation, Khmer uses the *flores200 tokenizer* (Costa-Jussà et al., 2022) and all the other languages uses the default *13a tokenizer* which is the sacreBLEU implementation of the mteval-13a tokenization script. All other settings were left to the defaults of sacreBLEU. Scores are only computed for datasets with available reference translations. Because the TORGO dataset only has the aligned text prompt and the corresponding audio file, BLEU cannot be calculated for the dysarthria dataset.

For both rhotacism and dysarthria, we use MetricX-24-Hybrid-XXL² and XCOMET-XL³ (Juraska et al., 2024; Guerreiro et al., 2024). These

²<https://huggingface.co/google/metricx-24-hybrid-xxl-v2p6>

³<https://huggingface.co/Unbabel/XCOMET-XL>

models estimate translation quality using the source and hypothesis alone and have been shown to correlate strongly with human judgments. MetricX⁴ emphasizes semantic adequacy, while XCOMET captures overall translation quality using cross-lingual contextual representations.

Together, these metrics provide a comprehensive evaluation of how speech impairments propagate through MT and how effectively input correction mitigates their impact.

6 Results & Discussion

Using the dysarthria and simulated rhotacism data described in Section 3, we first evaluate the effect of speech impairments and the ability of the three open-source LLMs to correct the errors. After this, we evaluate the downstream Machine Translation performance comparing the uncorrected to the corrected sentences.

All Word Error Rates (WER) are presented as percentages.

6.1 ASR Results

In Table 1, we show the WER for the dysarthric ASR output and the simulated rhotacism condition when compared to the correct reference sentences. The dysarthria transcript has a very large WER, above 100 indicating extremely poor ASR performance. Meanwhile the three rhotacism datasets achieve a small WER percentage appropriate for the proportion of words with /r/ sounds. These values act as baselines for our LLM correction experiments.

Dataset	WER ↓
Dysarthria	113.07
Rhotacism - Khmer	6.80
Rhotacism - Ukrainian	4.08
Rhotacism - Spanish	4.27

Table 1: Word Error Rates (WER) calculated against the correct transcriptions, showing authentic ASR performance for dysarthria alongside results from the three simulated rhotacism datasets.

⁴We use MetricX-24 rather than MetricX-25 because, although MetricX-25 has been described in the literature, the model was not publicly available at the time of our experiment Juraska et al. (2025).

6.2 LLM Correction Results

Table 2 presents the WER of each LLM-corrected output relative to the original ASR output, where a negative Δ indicates an improvement over the uncorrected baseline. For both dysarthria and rhotacism, model scale is a strong predictor of correction quality. Qwen3 (235B) achieves the best overall performance, attaining the lowest post-correction WER in nearly every setting and producing the largest reductions for both dysarthria ($\Delta = -1.40$) and the rhotacism conditions (Δ ranging from -2.47 to -2.73). LLaMA (70B) follows closely behind, also improving over the uncorrected baseline across all rhotacism datasets. OLMo (7B), by contrast, consistently increases WER relative to the uncorrected input, suggesting the model lacks sufficient capacity to reliably distinguish and repair impairment-induced errors.

Model	Dataset	WER ↓	Δ ↓
OLMo	Dysarthria	114.04	0.97
OLMo	Rhotacism - Khmer	9.64	2.84
OLMo	Rhotacism - Ukrainian	6.55	2.47
OLMo	Rhotacism - Spanish	7.74	3.47
LLaMA	Dysarthria	113.29	0.22
LLaMA	Rhotacism - Khmer	5.22	-1.58
LLaMA	Rhotacism - Ukrainian	3.05	-1.03
LLaMA	Rhotacism - Spanish	3.08	-1.19
Qwen3	Dysarthria	111.67	-1.40
Qwen3	Rhotacism - Khmer	4.07	-2.73
Qwen3	Rhotacism - Ukrainian	1.53	-2.55
Qwen3	Rhotacism - Spanish	1.80	-2.47

Table 2: WER of each LLM-corrected output compared to the correct transcriptions. Δ denotes the change relative to the uncorrected ASR baseline from Table 1, where negative values indicate improvement.

For dysarthria specifically, all three models struggle, with corrections yielding only marginal changes — even the best result from Qwen3 represents a modest 1.4% relative improvement. This is unsurprising given the severity of the dysarthria WER baseline (113.07), where the ASR output is heavily degraded and the correction signal is correspondingly weak. The rhotacism conditions, while

Experiment	Target Language	XCOMET $\uparrow$	MetricX $\downarrow$
Uncorrected	Khmer	0.6466	10.4437
OLMo Corrected	Khmer	0.6438	10.2847
LLaMA Corrected	Khmer	0.6518	10.1358
Qwen3 Corrected	Khmer	0.6512	10.1901
Clean	Khmer	0.8014	6.2781
Uncorrected	Ukrainian	0.7321	9.4699
OLMo Corrected	Ukrainian	0.7347	9.3444
LLaMA Corrected	Ukrainian	0.7414	9.1619
Qwen3 Corrected	Ukrainian	0.7388	9.2161
Clean	Ukrainian	0.9212	4.0137
Uncorrected	Spanish	0.7677	8.9878
OLMo Corrected	Spanish	0.7645	8.8321
LLaMA Corrected	Spanish	0.7716	8.3588
Qwen3 Corrected	Spanish	0.7739	8.4062
Clean	Spanish	0.9388	2.3965

Table 3: Downstream MT results for Dysarthria across three target languages. *Clean* denotes translation from ground-truth transcriptions; *Uncorrected* denotes translation from unmodified ASR output.

lower in overall error rate, show a clearer correction benefit, particularly for Qwen3 and LLaMA, whose post-correction WERs approach near-clean levels in several cases.

6.3 Downstream MT Results

Tables 3 and 4 present the downstream MT results for dysarthria and rhotacism respectively. In each table, the *clean* row represents the theoretical performance ceiling (translation from ground-truth transcriptions), while the *uncorrected* row represents the floor, reflecting translation directly from unmodified ASR output. The LLM-corrected rows fall between these two bounds and are the focus of this analysis.

Dysarthria. Despite the modest WER improvements observed in Table 2, LLM-based correction still yields consistent downstream translation gains for the two largest models. LLaMA achieves the best XCOMET and MetricX scores in two of the three language directions (Khmer and Ukrainian), while Qwen3 has the best XCOMET scores for Spanish. The two models are largely comparable, with differences falling within a narrow range. OLMo, consistent with its correction results, fails to improve over the uncorrected baseline XCOMET scores for Khmer and Spanish, though it does show minor MetricX gains across all three directions. Overall, improve-

ments remain modest relative to the gap between the uncorrected and clean conditions, which is expected given the limited correction gains at the ASR stage.

Rhotacism. The rhotacism results are more clear. Qwen3 generally yields the highest translation scores across conditions, though its advantage over LLaMA is typically modest and not consistent across all settings. Notably, Qwen3-corrected output approaches near-clean translation performance in several cases — for Ukrainian and Spanish in particular, the gap between Qwen3-corrected and clean narrows substantially across both XCOMET and MetricX. This is consistent with the near-zero post-correction WERs observed for Qwen3 in Table 2, suggesting that when surface-level errors are systematic and correctable, the effect of strong LLM correction can propagate meaningfully through the downstream pipeline. OLMo again lags behind, showing only marginal gains over the uncorrected baseline and failing to match LLaMA or Qwen3 in any direction.

6.4 Summary

Taken together, these results highlight two consistent themes. First, model scale matters: Qwen3 and LLaMA outperform OLMo at both the correction and translation stages across nearly every condition, while OLMo frequently fails to improve over — or

Experiment	Target Language	BLEU $\uparrow$	XCOMET $\uparrow$	MetricX $\downarrow$
Uncorrected	Khmer	8.76	0.8663	4.9233
OLMo Corrected	Khmer	8.74	0.8748	4.6139
LLaMA Corrected	Khmer	9.65	0.9046	3.8716
Qwen3 Corrected	Khmer	9.54	0.9122	3.6724
Clean	Khmer	9.71	0.9240	3.2711
Uncorrected	Ukrainian	34.00	0.9231	3.7415
OLMo Corrected	Ukrainian	34.42	0.9362	3.2167
LLaMA Corrected	Ukrainian	35.69	0.9544	2.6289
Qwen3 Corrected	Ukrainian	36.37	0.9584	2.4417
Clean	Ukrainian	37.55	0.9660	2.1251
Uncorrected	Spanish	47.43	0.9275	3.3913
OLMo Corrected	Spanish	48.07	0.9420	2.5573
LLaMA Corrected	Spanish	50.78	0.9565	1.9746
Qwen3 Corrected	Spanish	51.22	0.9632	1.4978
Clean	Spanish	52.73	0.9690	1.1615

Table 4: Downstream MT results for simulated rhotacism across three target languages. *Clean* denotes translation from ground-truth transcriptions; *Uncorrected* denotes translation from unmodified simulated ASR output.

actively degrades — the uncorrected baseline. Second, even modest surface-level corrections can yield measurable downstream benefits.

For dysarthria, where WER reductions are small, LLM correction still produces consistent translation gains for the two larger models. For rhotacism, where the errors are more systematic and more correctable, this effect is amplified — with Qwen3-corrected output approaching clean translation quality in the three language directions examined. These findings underscore both the promise and the current limitations of LLM-based ASR post-correction as a drop-in component for cascaded speech translation pipelines serving speakers with speech impairments.

Representative transcript examples illustrating both ASR errors and LLM corrections are provided in Appendix A.

7 Conclusion

In this work, we presented a systematic evaluation of speech impairment effects in cascaded speech-to-text translation, examining both real dysarthric speech and a text-based simulated rhotacism condition. We compared three open-source LLMs of varying scale as ASR post-correction tools and measured the downstream impact on translation quality across three lan-

guage directions.

Overall, our results demonstrate that while LLM-based correction yields only modest WER improvements—particularly for dysarthria—these surface-level gains can nonetheless produce consistent downstream translation benefits when model scale is sufficient. Qwen3 and LLaMA both improve over the uncorrected baseline across nearly all conditions, with Qwen3-corrected rhotacism output approaching clean translation quality in the three language directions examined. OLMo, at 7B parameters, lacks the capacity to reliably correct either condition. Taken together, these findings highlight both the potential and the current limitations of LLM post-correction as an accessible, training-free intervention for improving cascaded pipeline robustness for atypical speech.

Several directions remain open for future work. Most directly, our rhotacism experiments rely on simulated ASR errors rather than real recordings; collecting and evaluating on authentic rhotacism audio would provide a stronger empirical foundation. More broadly, extending this methodology to other speech disorders—such as stuttering or apraxia—would help characterize how impairment type and severity interact with downstream pipeline performance.

Thus, future work may benefit from speech-disability resources such as the Speech Accessibility Project, which contains substantially larger collections of authentic atypical speech Hasegawa-Johnson et al. (2024). Finally, it remains an open question whether other NLP technologies that consume spoken input, such as spoken dialogue systems or speech-based information retrieval, are similarly affected by impairment-induced ASR errors, and whether LLM post-correction offers comparable benefits in those settings.

Limitations

While our results offer meaningful insights into the effects of speech impairments on cascaded translation pipelines, several limitations should be noted.

Our rhotacism experiments rely on simulated ASR errors via minimal pair substitution rather than real speaker recordings, and future work should validate these findings using authentic rhotacism speech. The Torgo dataset also lacks reference translations, limiting MT evaluation to reference-free neural metrics (XCOMET and MetricX) and preventing direct comparison with rhotacism results. Because Tatoeba is publicly available, we cannot completely rule out the possibility of training-data overlap. We therefore interpret absolute translation scores cautiously and focus primarily on relative differences between conditions evaluated under identical settings. Beyond data constraints, our study examines only two speech impairments, and findings should not be assumed to generalize to conditions such as stuttering or apraxia of speech. The LLM correction experiments are similarly constrained by a single fixed prompt per condition; given known prompt sensitivity, our results likely represent a lower bound on what post-correction can achieve.

References

- Ahmed Aboeitta, Ahmed Sharshar, Youssef Nafea, and Shady Shehata. Bridging asr and llms for dysarthric speech recognition: Benchmarking self-supervised and generative approaches. *arXiv preprint arXiv:2508.08027*, 2025.
- Chitralkha Bhat and Helmer Strik. Two-stage data augmentation for improved asr performance for dysarthric speech. *Computers in Biology and Medicine*, 189: 109954, 2025.
- Carnegie Mellon University Speech Group. *The CMU Pronouncing Dictionary*. Carnegie Mellon University, 2014. Version 0.7b, available at <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.
- Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. Posix: A prompt sensitivity index for large language models, 2024. URL <https://arxiv.org/abs/2410.02185>.
- Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.
- Luigi De Russis and Fulvio Corno. On the impact of dysarthric speech on contemporary asr cloud platforms. *Journal of Reliable Intelligent Environments*, 5(3):163–172, 2019.
- M Dhanalakshmi, S Bharathipriya, E Vasantha Surya, and R Krishna Raj. Automatic speech recognition system for dysarthria leveraging mel spectrograms and transfer learning. In *2025 IEEE International Conference on Intelligent Signal Processing and Effective Communication Technologies (INSPECT)*, pages 1–6. IEEE, 2025.
- Matti Di Gangi, Robert Enyedi, Alessandra Brusadin, and Marcello Federico. Robust neural machine translation for clean and noisy speech transcripts. In Jan Niehues, Rolando Cattoni, Sebastian Stüker, Matteo Negri, Marco Turchi, Thanh-Le Ha, Elizabeth Salesky, Ramon Sanabria, Loic Barrault, Lucia Specia, and Marcello Federico, editors, *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong, November 2-3 2019. Association for Computational Linguistics. URL <https://aclanthology.org/2019.iwslt-1.32/>.
- Peter Flipsen Jr. *Remediation of/r/for speech-language pathologists*, volume 1. Plural Publishing, 2021.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. xcomet:

- Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024.
- Mark Hasegawa-Johnson, Xiuwen Zheng, Heejin Kim, Clarion Mendes, Meg Dickinson, Erik Hege, Chris Zwillig, Marie Moore Channell, Laura Mattie, Heather Hodges, et al. Community-supported shared infrastructure in support of speech accessibility. *Journal of Speech, Language, and Hearing Research*, 67(11):4162–4175, 2024.
- Yongjun He, Guanglu Sun, and Jiqing Han. Spectrum enhancement with sparse coding for robust speech recognition. *Digital Signal Processing*, 43:59–70, 2015.
- Abner Hernandez, Paula Andrea Pérez-Toro, Elmar Nöth, Juan Rafael Orozco-Arroyave, Andreas Maier, and Seung Hee Yang. Cross-lingual self-supervised speech representations for improved dysarthric speech recognition, 2022. URL <https://arxiv.org/abs/2204.01670>.
- Abner Hernandez, Tomás Arias-Vergara, Andreas Maier, and Paula Andrea Pérez-Toro. Enhancing asr accuracy for speakers with parkinson’s disease using instruction-tuned llms. In *International Conference on Text, Speech, and Dialogue*, pages 153–164. Springer, 2025a.
- Abner Hernandez, Tomás Arias Vergara, Andreas Maier, and Paula Andrea Pérez-Toro. Confidence-guided error correction for disordered speech recognition. *arXiv preprint arXiv:2509.25048*, 2025b.
- Yu-Chen Hsu and Yue-Shan Chang. A dysarthric speech recognition system with personal style embedding. In *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3770–3775. IEEE, 2025.
- Yuchen Hu, Chen Chen, Chao-Han Yang, Chengwei Qin, Pin-Yu Chen, Eng-Siong Chng, and Chao Zhang. Self-taught recognizer: Toward unsupervised adaptation for speech foundation models. *Advances in Neural Information Processing Systems*, 37:29566–29594, 2024.
- Shangkun Huang, Jing Deng, Jintao Kang, and Rong Zheng. Leveraging llm for stuttering speech: A unified architecture bridging recognition and event detection. *arXiv preprint arXiv:2505.22005*, 2025.
- Macarious Hui, Jinda Zhang, and Aanchan Mohan. Enhancing aac software for dysarthric speakers in e-health settings: an evaluation using torgo. In *ICC 2025-IEEE International Conference on Communications*, pages 3673–3679. IEEE, 2025.
- Artyom Iudin, Dmitrii Korzh, Matvey Skripkin, Dmitrii Masnyi, and Oleg Rogov. Enhancing speech recognition with llms in post-correction settings.
- Madeline Jefferson. Usability of automatic speech recognition systems for individuals with speech disorders: Past, present, future, and a proposed model. 2019.
- Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.35. URL <https://aclanthology.org/2024.wmt-1.35/>.
- Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. Metricx-25 and gemspaneval: Google translate submissions to the wmt25 evaluation shared task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 957–968, 2025.
- Gautam Krishna, Mason Carnahan, Shilpa Shamapant, Yashitha Surendranath, Saumya Jain, Arundhati Ghosh, Co Tran, Jose Del R Millan, and Ahmed H Tewfik. Brain signals to rescue aphasia, apraxia and dysarthria speech recognition. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 6008–6014. IEEE, 2021.
- Moreno La Quatra, Alkis Koudounas, Valerio Mario Salerno, and Sabato Marco Siniscalchi. Exploring generative error correction for dysarthric speech recognition. *arXiv preprint arXiv:2505.20163*, 2025.
- Wonjun Lee, San Kim, and Gary Geunbae Lee. Enhancing dialogue speech recognition with robust contextual awareness via noise representation learning. In Tatsuya Kawahara, Vera Demberg, Stefan Ultes, Koji Inoue, Shikib Mehri, David Howcroft, and Kazunori Komatani, editors, *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 333–343, Kyoto, Japan, September 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.sigdial-1.30. URL <https://aclanthology.org/2024.sigdial-1.30/>.

- Haoyang Li, Changsong Liu, Wei Rao, Hao Shi, Sakriani Sakti, and Eng Siong Chng. Training-free intelligibility-guided observation addition for noisy asr. *arXiv preprint arXiv:2602.20967*, 2026.
- Zhaolin Li, Yining Liu, Danni Liu, Tuan Nam Nguyen, Enes Yavuz Ugan, Tu Anh Dinh, Carlos Mullov, Alexander Waibel, and Jan Niehues. KIT’s low-resource speech translation systems for IWSLT2025: System enhancement with synthetic data and model regularization. In Elizabeth Salesky, Marcello Federico, and Antonis Anastasopoulos, editors, *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 212–221, Vienna, Austria (in-person and online), July 2025. Association for Computational Linguistics. ISBN 979-8-89176-272-5. doi: 10.18653/v1/2025.iwslt-1.20. URL <https://aclanthology.org/2025.iwslt-1.20/>.
- Ye Bhone Lin, Thura Aung, Ye Kyaw Thu, and Thazin Myint Oo. Asr error correction in low-resource burmese with alignment-enhanced transformers using phonetic features. In *2025 20th International Joint Symposium on Artificial Intelligence and Natural Language Processing (ISAI-NLP)*, pages 1–6. IEEE, 2025.
- Haowei Lou, Chengkai Huang, Hye-young Paik, Yongquan Hu, Aaron Quigley, Wen Hu, and Lina Yao. Speechagent: An end-to-end mobile infrastructure for speech impairment assistance. *arXiv preprint arXiv:2510.20113*, 2025.
- Rao Ma, Mengjie Qian, Mark Gales, and Kate Knill. Asr error correction using large language models. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.
- Hari Krishna Maganti and Marco Matassoni. Auditory processing-based features for improving speech recognition in adverse acoustic conditions. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):21, 2014.
- Anirudh Mani, Shruti Palaskar, Nimshi Venkat Meripo, Sandeep Konam, and Florian Metze. Asr error correction and domain adaptation using machine translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6344–6348. IEEE, 2020.
- Marco Marini, Nicola Vanello, and Luca Fanucci. Optimising speaker-dependent feature extraction parameters to improve automatic speech recognition performance for people with dysarthria. *Sensors*, 21(19):6460, 2021.
- Andres Marzal and Enrique Vidal. Computation of normalized edit distance and applications. *IEEE transactions on pattern analysis and machine intelligence*, 15(9): 926–932, 2002.
- E Matusov, S Kanthak, and Hermann Ney. On the integration of speech recognition and statistical machine translation. In *Proc. Interspeech 2005*, pages 3177–3180, 2005.
- Kinfe Tadesse Mengistu and Frank Rudzicz. Adapting acoustic and lexical models to dysarthric speech. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4924–4927. IEEE, 2011.
- Bernd T Meyer, Thomas Brand, and Birger Kollmeier. Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes. *The Journal of the Acoustical Society of America*, 129(1):388–403, 2011.
- Anna Min, Chenxu Hu, Yi Ren, and Hang Zhao. When end-to-end is overkill: Rethinking cascaded speech-to-text translation. *arXiv preprint arXiv:2502.00377*, 2025.
- Megan A Morris, Sarah K Meier, Joan M Griffin, Megan E Branda, and Sean M Phelan. Prevalence and etiologies of adult communication disabilities in the united states: Results from the 2012 national health interview survey. *Disability and health journal*, 9(1):140–144, 2016.
- Denis Moser, Nikola Stanic, and Murat Sariyar. Benchmarking speech-to-text robustness in noisy emergency medical dialogues: an evaluation of models under realistic acoustic conditions. *JAMIA open*, 8(6):oofaf147, 2025.
- Ramy Mounir, Redwan Alqasemi, and Rajiv Dubey. Speech assistance for persons with speech impediments using artificial neural networks. In *ASME International Mechanical Engineering Congress and Exposition*, volume 58363, page V003T04A056. American Society of Mechanical Engineers, 2017.
- Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch,

- Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL <https://arxiv.org/abs/2512.13961>.
- Matt Post. A call for clarity in reporting bleu scores. In *Proceedings of the third conference on machine translation: Research papers*, pages 186–191, 2018.
- Jie Pu, Thai-Son Nguyen, and Sebastian Stüker. Multi-stage large language model correction for speech recognition. *arXiv preprint arXiv:2310.11532*, 2023.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.
- Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. The toro database of acoustic and articulatory speech from speakers with dysarthria. *Language resources and evaluation*, 46(4):523–541, 2012.
- Megha Rughani and D. Shivakrishna. Hybridized feature extraction and acoustic modelling approach for dysarthric speech recognition, 2015. URL <https://arxiv.org/abs/1506.02170>.
- Paban Sapkota, Hemant Kumar Kathania, Shrikanth Narayanan, and Sudarsana Reddy Kadiri. Addressing dysarthric speech variability with severity-aware fine-tuning of transformer-based wav2vec2 asr models: P. sapkota, hk kathania, s. narayanan, sr kadiri. *Circuits, Systems, and Signal Processing*, pages 1–41, 2026.
- Seyed Reza Shahamiri, Vanshika Lal, and Dhvani Shah. Dysarthric speech transformer: A sequence-to-sequence dysarthric speech recognition system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3407–3416, 2023.
- Mohammed Sidi Yakoub, Sid-ahmed Selouani, Brahim-Fares Zaidi, and Asma Bouchair. Improving dysarthric speech recognition using empirical mode decomposition and convolutional neural network. *EURASIP Journal on Audio, Speech, and Music Processing*, 2020(1):1, 2020.
- Zhiyuan Tang, Dong Wang, Shen Huang, and Shidong Shang. Full-text error correction for chinese speech recognition with large language model. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.
- Jörg Tiedemann. The tatoeba translation challenge—realistic data sets for low resource and multilingual mt. *arXiv preprint arXiv:2010.06354*, 2020.
- Jimmy Tobin, Phillip Nelson, Bob MacDonald, Rus Heywood, Richard Cave, Katie Seaver, Antoine Desjardins, Pan-Pan Jiang, and Jordan R Green. Automatic speech recognition of conversational speech in individuals with disordered speech. *Journal of Speech, Language, and Hearing Research*, 67(11):4176–4185, 2024.
- Zhong-Qiu Wang, Peidong Wang, and DeLiang Wang. Complex spectral mapping for single-and multi-channel speech enhancement and robust asr. *IEEE/ACM transactions on audio, speech, and language processing*, 28: 1778–1787, 2020.
- Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *International conference on latent variable analysis and signal separation*, pages 91–99. Springer, 2015.
- Cihan Xiao, Matthew Wiesner, Debashish Chakraborty, Reno Kriz, Keith Cunningham, Kenton Murray, Kevin Duh, Luis Tavarez-Arce, Paul McNamee, and Sanjeev Khudanpur. Whisper-ut: A unified translation framework for speech and text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3000–3016, 2025.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chen-

gen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. Generative speech recognition error correction with large language models and task-activating prompting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

Xiaohu Zhao, Haoran Sun, Yikun Lei, and Deyi Xiong. Regularizing cross-attention learning for end-to-end speech translation with asr and mt attention matrices. *Expert Systems with Applications*, 247:123241, 2024.

A Appendix

A.1 Prompt Template

Prompt for Corrections

This is a sentence in English that was transcribed by an ASR system. The person who spoke it has a speech impairment ([INSERT IMPAIRMENT TYPE]). Fix any errors caused by the speech impairment and ASR. If the sentence is not in English, correct it to proper English based on how it sounds. If the sentence appears to be accurate, return it as is with nothing extra. Return only the corrected sentence.
The sentence: [INSERT SENTENCE]

A.2 Examples from Transcripts

Rhotacism Example 1 (Line 43)

English

Original: Fresh fruit is good for you.

Uncorrected: flesh flute is good fall you.

OLMo Corrected: flesh flute is good for you.

LLaMA Corrected: Fresh fruit is good for you.

Qwen3 Corrected: Fresh fruit is good for you.

Spanish Translation

Original: Las frutas frescas son buenas para ti.

Uncorrected: La flauta de carne es buena para ti.

OLMo Corrected: La flauta de carne es buena para ti.

LLaMA Corrected: Las frutas frescas son buenas para ti.

Qwen3 Corrected: Las frutas frescas son buenas para ti.

Rhotacism Example 2 (Line 137)

English

Original: I am writing to express my dissatisfaction.

Uncorrected: I am lighting to express my dissatisfaction.

OLMo Corrected: I am lighting out to express my dissatisfaction.

LLaMA Corrected: I am writing to express my dissatisfaction.

Qwen3 Corrected: I am writing to express my dissatisfaction.

Spanish Translation

Original: Estoy escribiendo para expresar mi insatisfacción.

Uncorrected: Estoy encendida para expresar mi insatisfacción.

OLMo Corrected: Estoy encendida para expresar mi insatisfacción.

LLaMA Corrected: Estoy escribiendo para expresar mi insatisfacción.

Qwen3 Corrected: Estoy escribiendo para expresar mi insatisfacción.

LLMs as Translator Training Partners: A Multi-Agent Approach

Ming Qian

Human-Centered AI, Charles River Analytics, Cambridge, MA, USA

qianmi@gmail.com

Luyi Yang

Certified Translator and Interpreter, Toronto, ON, Canada

luyiyang335@gmail.com

Abstract

Peer-review-based translator training promotes reflection and collaborative critique. This study investigates whether GPT-5.4, guided by Multidimensional Quality Metrics (MQM) style prompts, can function as a peer-review training partner rather than a grading tool. Using translated passages from a translator practice group, we compared GPT-5.4's feedback with human peer-review comments on the same translation segments, including both negative and positive judgments, and used a second large language model (DeepSeek) to arbitrate disagreements between human and AI judgments. GPT-5.4 agreed with human trainees on the existence of translation issues in 77.8% of negative flags and 88.9% of positive flags and achieved an F1 score of 0.875 against in-depth verification by two human evaluators. Although GPT-5.4 occasionally produced incorrect judgments, it also generated more comprehensive and analytical feedback than human trainees and frequently offered alternative but equally valid interpretations that could stimulate deeper reflection. DeepSeek successfully resolved a substantial proportion of human-LLM disagreements, demonstrating the value of an independent LLM arbitrator. These findings suggest that a multi-agent framework combining an MQM-guided reviewer, human deliberation, and LLM-based arbitration is a feasible and promising approach for developing reliable AI training partners for translator education.

1 Introduction

Peer-review-based learning asks learners to evaluate and comment on one another's work. By comparing peers' feedback, learners engage in metacognitive reflection: examining alternative strategies, identifying strengths and weaknesses, and reconsidering their own choices. Exposure to diverse perspectives can further improve learning outcomes and strengthen group problem solving.

LLM-based quality checks, or "LLM-as-judge" systems, are fast and scalable but still struggle to capture nuance, address critical issues, and detect subtle domain-specific errors (Li et al., 2024; Wang, R., et al., 2025). Bavaresco et al. (2025) questioned their validity as substitutes for human judges in translation and NLP tasks, citing biases and inconsistencies with human assessments. Wang, G., et al. (2025) proposed a multi-agent automated translation system using the Multidimensional Quality Metrics (MQM) framework for fine-grained error detection and refinement but did not report empirical results. Overall, available evidence (Feng et al., 2025; Gu et al., 2024; Szymanski et al., 2025; Ye et al., 2024) suggests that LLM-as-judge systems are best used as auxiliary quality-control tools, not replacements for expert human evaluators, especially in high-stakes contexts such as ATA exam grading.

In this study, we examine LLMs as training partner

need training partners that are readily available, can identify problematic text, and provide feedback that encourages deeper reflection, even if they do not match grader-level accuracy. The core argument is that a training partner is valuable not because they are always correct, but because they create productive friction. If they sometimes challenge you, disagree with you, or even make occasional mistakes, they can still help you learn by forcing you to think, test your assumptions, and externalize your reasoning more clearly.

LLMs are well suited to this role because they are consistently accessible and offer a low-stress environment for practice. We used ChatGPT5.4 to generate peer-review feedback based on the MQM framework. MQM assesses translation quality through customizable error categories and severity levels, including accuracy, fluency, terminology, and style (Lommel et al., 2014; MQM Council, 2025). Prior work has used multi-agent LLM systems to automatically evaluate distinct MQM categories (Fernandes et al., 2023; Lu et al., 2025; Wang G. et al., 2025).

To evaluate our approach, we compared GPT-5.4-generated MQM evaluations with human trainees' peer-review comments on the same human-translated passage. The human review data came from a peer-led translation practice group organized by the ATA Chinese

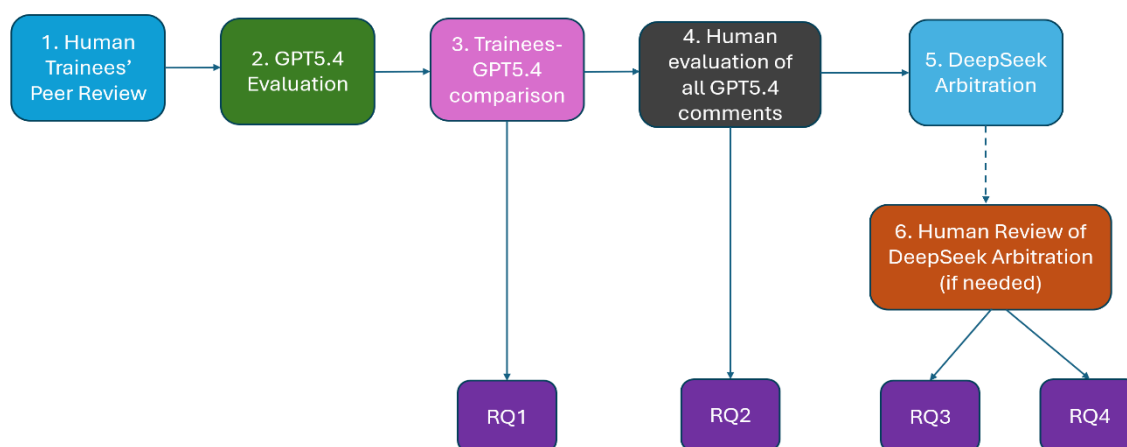


Figure 1: Experimental workflow and analysis process.

(American Translators Association – Community Led Division, n.d.). Four trainees translated a short exam-style passage and reviewed anonymized peer translations in a shared Google Doc. Trainees provided both positive and negative review comments. Notably, one trainee had passed the ATA certification exam before the practice in December 2023 and two trainees passed the ATA certification exam sometime after the practice.

We then prompted GPT-5.4 using an MQM-style prompt to generate positive and negative review comments for the same passage. First, our analysis focused on translation segments that were annotated by both the human trainees and GPT-5.4. Specifically, we addressed two research questions:

1. What percentage of the human trainees' positive and negative comments were also identified by GPT-5.4? How do comments with the same overall polarity differ in error types and severity?
2. Since GPT-5.4 generated substantially more comments than the human trainees, two human evaluators assessed all GPT5.4 comments to determine what percentage of its positive and negative comments could be confirmed as valid (we evaluate both true and false positives and negatives and calculate the F1 score as a quality indicator).

A second large language model (DeepSeek-V4) was used to arbitrate cases in which GPT-5.4 disagreed with the human trainees and/or evaluators. For cases where DeepSeek's judgments conflicted with the evaluators' assessments, the two evaluators conducted detailed reviews to reach a final consensus on the correctness of DeepSeek's arbitration. We investigated the following research questions:

3. What percentage of DeepSeek's arbitration decisions were judged to be correct?
4. How much does the human-LLM deliberation workflow, followed by LLM

process?

Figure 1 illustrates the experimental workflow and analysis process.

2 Experimental Data and MQM-like Prompts

Four trainees independently translated the same English-to-Chinese passage and evaluated each other's work. MQM-style prompts were then used to generate GPT5.4 evaluation comments. Because GPT5.4 produced far more comments than the human trainees, two evaluators reviewed all GPT5.4 comments to validate them. Implementation details follow.

2.1 Experimental data

Four trainees were each assigned to translate the same English passage—comparable in length and difficulty to those on the ATA certification exam—into Chinese. All translations were submitted anonymously in a shared Google Doc, after which each trainee reviewed the others' work. *Below is the English passage (an excerpt from: <https://www.mckinsey.com/featured-insights/future-of-asia/videos/is-asia-on-the-cusp-of-a-new-era>):*

“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years. Asia will be at the nexus of these forces and will experience heightened versions of the trends affecting the rest of the world.

Asia is the world's trade crossroads. It's home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest. But now, it could find itself in the crosshairs of trade tensions. Asia's integration has come about through commercial pragmatism rather than political alignment. Can that continue if trade tensions rise?

Asia excels in manufacturing new technology in four areas in particular: consumer electronics, automobiles, and

semiconductors. In these places, Asia has more than 40 percent of the revenues, R&D, and patents. But the value created by tech is moving away from hardware and into new transversal technologies. So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.

Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies. But now, aging is fiercest in the high-productivity economies of the Pacific Rim. Asia has the people to fuel growth, but 90 percent of the expected increase in Asia's non-farm workforces by 2050 will be where productivity is relatively low. Can Asia shift value chains to where the workers are and raise productivity everywhere?"

2.2 MQM-like Prompts

For GPT5.4-generated negative comments, we used the following prompt based on the prompt proposed in Wang et al. (2025):

"The whole source English text segment: ...

The source text sentence:

"We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years."

The translated target text sentence:

"我们相信新的时代马上就要来临，这个新时代深层暗涌的力量与在过去30年促成世界格局的力量大不相同。"

You are a professional translator responsible for evaluating the quality of the translated target text sentence based on the Multidimensional Quality Metrics (MQM) framework.

Your task is to identify all errors in the translation and classify them into the following dimensions:

Accuracy (Mistranslation, omission, addition, over-translation),

Terminology (Incorrect term choice, missing glossary application),

Linguistic Conventions (Grammar mistakes, punctuation errors, typographical issues),

Style (Unnatural flow, awkward expressions, overly literal phrasing),

Locale Conventions (Wrong format, incorrect symbols),

Audience Appropriateness (Overly technical terms for general audiences, offensive wording),

Design and Markup (Missing markup, broken text formatting, layout inconsistencies),

Each error must

Critical: Critical errors inhibit comprehension of the text.

Major: Major errors disrupt the flow, but what the text is trying to say is still understandable.

Minor: Minor errors are technically errors, but do not disrupt the flow or hinder comprehension.

If no error is detected, return "no-error" in its severity."

For GPT5.4-generated positive comments, we used the following prompt:

"The whole source English text segment: ...

The source text sentence:

"We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years."

The translated target text sentence:

"我们相信新的时代马上就要来临，这个新时代深层暗涌的力量与在过去30年促成世界格局的力量大不相同。"

You are a professional translator responsible for evaluating the quality of the translated target text sentence based on the Multidimensional Quality Metrics (MQM) framework.

Please give some feedback on where the translation is done exceptionally well and commendable:

Accuracy (meaning is accurately conveyed across language and culture, no mistranslation, no omission, no addition, no over-translation),

Terminology (correct term choice),

Linguistic Conventions (no grammar mistakes, no punctuation errors, no typographical issues),

Style (flows eloquently and naturally as if a native speaker who has a good command of the target language wrote it; no awkward expressions and not read like a rigid translation, no overly literal phrasing)

Locale Conventions (correct format, correct symbols),

Audience Appropriateness (proper technical terms for educated laymen, no offensive wording).

If none is detected as exceptionally well, return "Nothing is found."

..

It's important to note that, before the prompts, we include the full source paragraph, the specific source sentence of interest, and the translated target sentence that is under evaluation.

2.3 DeepSeek Arbitration Prompt

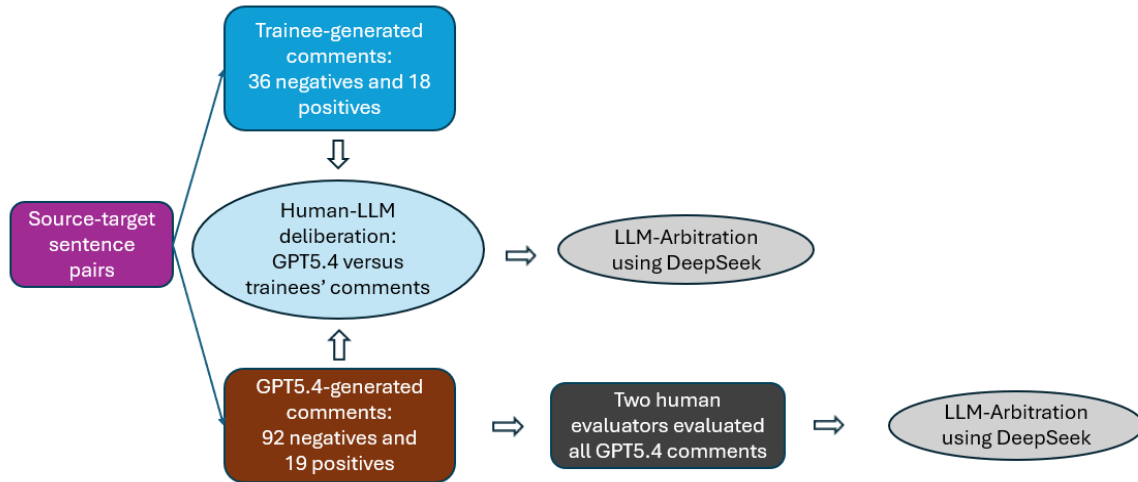


Figure 2: Collecting performance statistics based on trainee-generated and GPT5.4-generated comments through human-LLM deliberation and LLM-arbitration.

We used the following prompt to elicit DeepSeek arbitration:

“There are two comments on the translation pair below:

“Asia has more than 40 percent of the revenues”

“亚洲占了超过百分之四十的收益”

Comment 1: “40% “ should be written in Chinese text ” 百分之四十 “. Are there any specific formatting rules?

Comment 2: No problem detected.

Which comment is better? Why?

”

2.4 Evaluations

We established a human-generated baseline by collecting four trainees’ translation outcomes and their review comments towards each other. Altogether, we collected 32 negative, 18 positive trainee-generated comments, and the corresponding source/target sentence pairs.

Using the collected source–target sentence pairs and the MQM-style prompts described in Section 2.2, we generated both positive and negative comments with GPT5.4. Based on these results, we computed the following statistics:

- (1) What percentage of trainee-generated negative comments were flagged by GPT5.4 using the MQM-style prompt?
- (2) What percentage of trainee-generated positive comments were flagged by GPT5.4 using the MQM-style prompt?
- (3) What percentage of GPT-generated negative comments were confirmed by the two human evaluators?
- (4) What percentage of GPT-generated positive comments were confirmed by the two human evaluators?

DeepSeek (a second LLM) was used to arbitrate cases in which GPT-5.4 disagreed with the human trainees and/or evaluators. For cases where DeepSeek’s judgments conflicted with the evaluators’ assessments, the two evaluators conducted (additional) detailed reviews to reach a final consensus on the correctness of DeepSeek’s arbitration. Based on these results, we computed the following statistics

- (5) What percentage of DeepSeek’s arbitration decisions were judged to be correct?
- (6) How much does the human–AI deliberation workflow, followed by LLM arbitration, improve the overall evaluation process?

In terms of performance measures:

- (1) We calculated the accuracy and misclassification rates of GPT5.4 in confirming human-generated negative and positive comments.
- (2) For GPT5.4-generated negative and positive comments, we calculated true positives, true negatives, false positives, false negatives, and the F1 score based on the two evaluators’ investigation.

Figure 2 illustrates how performance statistics are collected based on trainee-generated and GPT5.4-generated comments through human-LLM deliberation and LLM-arbitration.

3 Results and Discussion

We measured both GPT5.4 accuracy relative to trainees’ judgments (using trainee-identified negatives/positives as the baseline) and classification performance of GPT5.4-generated comments (using GPT5.4-generated negatives/positives as the baseline).

3.1 GPT5.4 Accuracy Relative to Trainees’ Judgments

Appendix Table 1 summarizes the 36 errors

error existence was treated as a binary decision. Human trainees and GPT-5.4 could agree that an error existed while disagreeing on its type or severity. GPT-5.4 agreed with the trainees on error existence in 28 of the 36 cases. Among these shared errors, they agreed on the error type in 17 of 28 cases and on error severity in 18 of 28 cases. In the remaining 8 cases, GPT-5.4 did not identify an error that the human trainees had flagged. The two human evaluators determined that these disagreements primarily reflected trainee nitpicking of very minor issues, suggesting that GPT-5.4 reduced over-flagging of trivial issues. However, the evaluators also found that GPT-5.4 overstated error severity in 9 cases by classifying minor errors as major, indicating a tendency to overestimate severity after the error existence was established.

In four instances, GPT5.4 and human trainees marked the same text span as erroneous but assigned different error types. In text segment 8, a human trainee labeled “生产力最高的经济体” as a mistranslation of “high-productivity economies,” because “最高” means “highest,” whereas the source indicates “high,” is not a superlative. GPT5.4 instead labeled it as a style error, noting that the rendering was overly literal and less natural than standard Chinese economic terminology. In this case, the trainee’s annotation is more accurate. Also in text segment 8, a human trainee flagged “老龄化……最为激烈” as an unnatural expression, while GPT5.4 classified it as a word-choice nuance error. In this case, neither annotation is clearly preferable. In text segment 9, a human trainee found that “fuel growth” was mistranslated, whereas GPT5.4 described the translation as overly literal. The human trainee’s annotation is more accurate. In text segment 13, a human annotator flagged an incorrect translation of “surging,” while GPT5.4 classified it as an awkward translation. Here again, the trainee’s annotation provides the more accurate explanation. Overall, when human trainees and GPT5.4 disagree on error type, the human annotations are generally more accurate and address root causes instead of surface effects.

Appendix Table 2 lists 18 cases where human trainees identified notable strengths rather than errors. When evaluating the same sentence pairs, GPT5.4 agreed in 16 of 18 cases. In the remaining two cases, GPT5.4 judged the translations as merely acceptable, whereas trainees found them to be exceptional—the two evaluators determined that the two cases were not strong at all. Humans often elaborate more on criticism than praise because of negativity bias (“Negativity Bias,” 2026). In actual study groups, human trainees’ praise is often shorter and more judicious.

pre-designed feedback typically requested in controlled experiments. In contrast, GPT5.4, when properly prompted, tends to give more analytical, evaluative, and detailed explanations for positive feedback. For example,

Trainee’s positive comment: “新时代的风口浪尖” is a vivid translation.

GPT5.4: The translation, “新时代的风口浪尖”, adds rhetorical strength and make the sentence vivid, while still staying close to the original meaning. It does not feel mechanical or overly literal.

3.2 Classification Performance of GPT5.4-generated comments

For the same 28 sentence pairs that human translators marked with 36 negative or 18 positive comments, GPT5.4 generated 92 negative comments and 19 positive comments. Appendix Table 3 lists the 92 cases in which GPT5.4 flagged errors; the two authors (evaluators) evaluated these cases and found that 77 were reasonable and 15 were false. Appendix Table 4 lists the 19 cases in which GPT5.4 noted strengths; the two evaluators found that 12 were reasonable and 7 were false.

To evaluate GPT5.4 as an AI system for automatically commenting on translation quality, the F1 score is calculated, based on true positives, false positives, true negatives, and false negatives, as $(2 \times 77) / (2 \times 77 + 15 + 7) = 0.875$. There is no universal “good” F1-score; it depends on the task and the cost of false positives versus false negatives. In practice, 0.8 to 0.9 is often considered good, and above 0.9 is usually excellent. Therefore, GPT5.4, prompted by the MQM-like prompt can be considered a good AI system for this small task in this study.

Below is a list of reasons we identified where GPT5.4 generated the false negatives (examples in which GPT5.4 identified as false while human evaluators deemed acceptable):

- Misclassifying a suboptimal but acceptable translation as unacceptable.
- Incorrectly flagging meaning omissions or additions: translation is not always a one-to-one mapping. A source sentence may be split into multiple target sentences, or several source clauses may be combined into one coherent translated sentence.
- False fault isolation: mistranslation may stem from a misunderstanding of the source sentence as a whole but GPT5.4 may incorrectly identify the translation of a specific source word as the cause of the problem instead of a sentence-level issue.

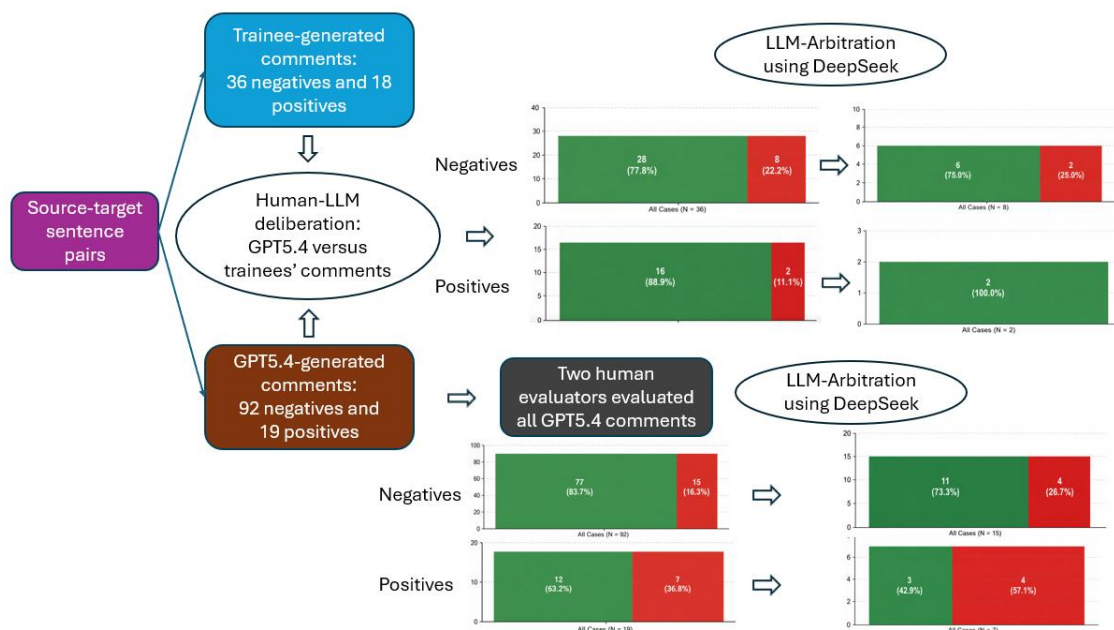


Figure 3: Comparison of LLM deliberation and arbitration results with those of human trainees and evaluators.

ignore the fact that syntax works differently across languages and expect parts of speech in the target text to align perfectly with the source, even when the same meaning is expressed naturally through different structures.

- Overly strict structural matching: GPT5.4 may complain that the target expression introduces an structure not present in the source sentence and treat it as a mistranslation, even when that structure makes the translation more natural in the target language.

- Overreliance on common translation choices: GPT5.4 may rigidly insist on the most commonly seen translation and ignore contextual factors that could justify alternative choices.

- Overly narrow word interpretation: GPT5.4 may narrowly limit a word's meaning or function in the target language and ignore other valid meanings or alternative functions the word can serve.

- Inconsistency: Inconsistent error detection: GPT5.4 may flag a repeated translation phrase as a mistranslation in one instance while ignoring the same issue in other occurrences.

- Ignoring contextual meaning: GPT5.4 may criticize the translation for adding meaning, even when that meaning is already implied or established by the surrounding context.

Below is a list of reasons we identified where GPT5.4 generated the false positives (examples in which GPT5 noted positive strength while human evaluators disagree):

- Over-praising weak word choices: GPT5.4 may praise a suboptimal word choice even when it dist

- Over-praising over-translation: GPT5.4 may praise an over-translated phrase even when it adds meaning, emphasis, or specificity that were not present in the source.
- Over-praising fancy translation choices: GPT5.4 may praise a polished or sophisticated translation choice even when it introduces added inaccuracy or distorts the source meaning.
- Over-praising poorly adopted word choices: GPT5.4 may praise a word choice that is not well adopted or natural in the target language, even when it sounds awkward or non-idiomatic.
- Over-praising idiomatic but inaccurate translations: GPT5.4 may praise an idiomatic translation even when it distorts the source meaning.

GPT5.4 appears to evaluate each translation segment across all error categories (that could be a strength), sometimes forcing comments even when a category is not relevant. Human trainees, by contrast, typically identify the most salient error category and leave unnecessary categories unaddressed (saliency bias). When an error could fall under multiple categories, humans focus on the root cause. Criticizing a single sentence across many categories can be overly harsh and unnecessary; it is unlikely that one sentence would contain every possible type of translation error.

In contrast, GPT-5.4 can analyze subtle and nuanced cases consistently and in detail without experiencing the fatigue that may affect

3.3 Arbitration Performance of DeepSeek

DeepSeek-V4 was used to arbitrate disagreements between GPT-5.4 and the human trainees. Appendix Table 5 summarizes DeepSeek's arbitration of disagreements on negative exemplars. After further review, the human evaluators determined that DeepSeek's decisions were correct in 6 of the 8 cases. Appendix Table 6 summarizes DeepSeek's arbitration of disagreements on positive exemplars. The human evaluators confirmed that DeepSeek's decisions were correct in both cases (2 of 2).

DeepSeek was also used to arbitrate disagreements between GPT5.4 and the two human evaluators. Appendix Table 7 lists DeepSeek arbitration towards disagreements on negative exemplars between human evaluators and GPT5.4. Based on the human evaluators' further investigation and final judgments, DeepSeek's arbitration was correct in 11 of the 15 cases. Appendix Table 8 lists DeepSeek arbitration towards disagreements on positive examples between human trainees and GPT5.4. Based on the human evaluators' further investigation and final judgments, DeepSeek's arbitration was correct in 3 of the 7 cases.

Figure 3 shows comparison of LLM deliberation and arbitration results with those of human trainees and evaluators.

Discussion and Conclusion

Melby (2012) defined TTT (Translation Turing Tests or Translation Technology Tests) using an operational approach: human, machine, or hybrid translation providers are expected to generate translation results that meet detailed specifications. ATA exam passages are typically written at “educated layperson level” and comes with a brief instruction of the intended audience (e.g., “for university students learning about the subject”). Other than a general readability level and a brief indication of the intended audience, the specification of ATA exam passages is relatively general.

General specifications allow greater room for human judgment. Unlike human graders in translation certification exams such as the ATA exam, who focus on formal assessment, human trainees may be less accurate but can still provide valuable opportunities for learning and practice. Because we compared GPT-5.4 with human trainees rather than professional graders, the human benchmark represents a less stringent standard. Therefore, the Turing test posed to GPT-5.4 is relatively easier to pass.

What we need most from training partners is availability, attention, the ability to be a “more knowledgeable other”, and productive friction that encourages critical, creative, and deeper reflection.

available and can provide a low-stress practice environment, since learners may feel less fear of embarrassment or negative judgment even when making serious mistakes. In addition, unlike human partners, LLMs would never experience fatigue.

We identify the following strengths of using LLMs as a training partner (guided by an MQM-style prompt):

- It can identify most major/critical errors identified by human trainees.
- It can reduce over-flagging of trivial (nitpicking) issues.
- It can detect and document a greater number of errors (without fatigue) and achieve good performance (F1 score = 0.875 based on our small experiment).
- With appropriate prompting, it can provide reasons and explanations for positive cases.
- It could generate alternative error types that encourage reflection and the human-LLM team to reach a more balanced judgment.

We identify the following weaknesses of using LLMs as a training partner (guided by an MQM-style prompt):

- It can overstate/inflate error severity.
- It can generate less accurate error types.
- It can generate a greater number of errors without proper saliency bias (identify the most salient error category while leave out secondary issues)
- LLMs, like human, can make poor judgment. The issues included misclassifying suboptimal but acceptable translations as unacceptable; incorrectly flagging meaning omissions or additions; attributing errors to the wrong linguistic element; requiring overly rigid syntax or structural alignment between source and target; relying too heavily on common translation choices or narrow word interpretations; applying inconsistent standards; ignoring contextual meaning; and over-praising weak, overly literal, overly embellished, or poorly adapted word choices, even when they introduced inaccuracies.

DeepSeek was used to arbitrate disagreements regarding the existence of translation errors between GPT-5.4 and both the human trainees and the human evaluators. Our findings suggest that incorporating a second LLM as an independent arbitrator can help resolve a substantial proportion of human-LLM disagreements, providing an additional layer of

assisted translator training.

In a peer-learning setting, stronger trainees can serve as a more knowledgeable other, helping their peers while also reinforcing their own learning by explaining their reasoning, answering questions, and comparing approaches. In our experiment, GPT-5.4 demonstrated the characteristics of a more knowledgeable other by identifying most of the errors detected by human trainees, as well as a greater number of subtle and nuanced errors. It also maintained a high F1 score, indicating both high precision and a low false negative rate. However, we would not conclude that GPT-5.4 definitively passes the translation Turing test. A Turing test is not simply about achieving a high F1 score; it asks whether humans can reliably distinguish GPT-5.4 from human trainees/evaluators based on relevant behaviors. Although we observed substantial behavioral similarity, we cannot establish indistinguishability.

Although our study was conducted on a relatively small dataset, the results provide encouraging evidence that a multi-agent approach to developing LLM-based training partners for translators is both feasible and promising. The combination of an MQM-guided reviewer, human deliberation, and a second LLM acting as an independent arbitrator demonstrates the potential to provide reliable feedback, encourage critical reflection, and improve judgment quality. Future work should validate this approach on larger datasets, additional language pairs, and more diverse translation domains to further assess its effectiveness and generalizability.

Acknowledgements

We sincerely thank Roanne Cheung, Jessie Doherty, and Zhitong Zou from the ATA-CLD translation practice group for their dedicated efforts in generating high-quality annotations and generously sharing their results, which were essential to this study's dataset.

References

American Translators Association – Community-Led Division. (n.d.). *Translation Practice Group*. ATA Divisions. <https://www.ata-divisions.org/CLD/resources/translation-practice-group/>

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., ... & Testoni, A. (2025, July).

Proceedings of the 17th Conference of the Association for Machine Translation in the Americas, Llm instead of hu Québec City, Canada, August 31 - September 2, 2026. Volume 1: Research Track

study across 20 nlp evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 238-255).

Feng, Z., Su, J., Zheng, J., Ren, J., Zhang, Y., Wu, J., ... & Liu, Z. (2025, July). M-MAD: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 7084-7107).

Fernandes, P., Deutsch, D., Finkelstein, M., Riley, P., Martins, A. F., Neubig, G., ... & Firat, O. (2023, December). 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* (pp. 1066-1083).

Kocmi, T., & Federmann, C. (2023, December). GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 768-775).

Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., ... & Liu, Y. (2024). Llm-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumática*, (12), 0455-463.

Lu, Q., Ding, L., Zhang, K., Zhang, J., & Tao, D. (2025, January). 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*

Melby, A. (2012). Human and machine translation quality: Definable? Achievable? Desirable. In LACUS Forum (Vol. 39, pp. 1-29). Available online: www.ttt.org/melbyak [last accessed 18 August. 2026]

MQM Council. (2025). *MQM (Multidimensional Quality Metrics)*. <https://themqm.org/>

Negativity bias (2026, April 30th). In Wikipedia https://en.wikipedia.org/wiki/Negativity_bias

Szymanski, A., Ziems, N., Eicher-Miller, H. A., Li, T. J. J., Jiang, M., & Metoyer, R. A. (2025, March). Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th international conference on intelligent user interfaces* (pp. 952-966).

Wang, G., Hu, J., & Ali, S. (2025). MAATS: A Multi-Agent Automated Translation System Based on MQM Evaluation. *arXiv preprint arXiv:2505.14848*.

Wang, R., Guo, J., Gao, C., Fan, G., Chong, C. Y., & Xia, X. (2025). Can llms replace human evaluators? an empirical study of llm-as-a-judge in software engineering. *Proceedings of the ACM on Software Engineering*, 2(ISSTA), 1955-1977.

Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., ... & Zhang, X. (2024). Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Appendix

#	Source and Target Sentences	Human & GPT5 generated feedback
1	“We believe that <i>the world</i> is on the cusp of a new era...” “我们相信新的时代马上就要来临...”	Trainee: “the world” was not translated. (Severity: Minor) GPT5: omission (Severity: Minor) Same error type
2	“But now, it could find itself in the crosshairs of trade tensions.” “当今贸易 紧张局势 也在这里得到了格外强烈的 体现”	Trainee: “crosshairs of trade tensions: ➔ “紧张局势.....的体现” is a mis-translation. (Severity: Minor) GPT5: mis-translation (Severity: Major) Same error type
3	“Asia’s integration has come about through commercial pragmatism rather than political alignment.” “亚洲的融入是通过商业务实性，而不是政治 取向 一致性。”	Trainee: The phrase “取向” is a mis-translation. (Severity: Minor) GPT5: Incorrect term choice (Severity: Major) Both error types are acceptable. Mistranslation is “meaning error” while the term choice “取向” is incorrect.
4	“Can that continue if trade tensions rise?” “亚洲还可以维持现状吗?”	Trainee: "That" in the original text refers more to the strategy of commercial pragmatism, so it doesn't mean “maintaining the status quo.” (“维持现状”) (Severity: Major) GPT5: Mis-translation (Severity: Major)

meaning error.

- 5 “Asia excels in manufacturing new technology in four areas in particular”
“亚洲在如下四项新型科技制造方面超群拔类”
Trainee: The original text only discussed technology; translating it as “科技” adds meaning not present in the original text. (Severity: Minor)
GPT5: incorrect term choice; vague and non-standard (Severity: Major)
Both error types are acceptable: minor addition versus incorrect term choice.
- 6 “consumer electronics, industrial electronics, electric vehicles, and semiconductors.”
“消费电子产品、工业电子产品、电动汽车、半导体。”
Trainee: The last item in a list must be connected with a “和”. (Severity: No error)
GPT5: No problem detected
- 7 “Asia has more than 40 percent of the revenues”
“亚洲占了超过百分之四十的收益”
Trainee: Why “40% “ should be written in Chinese text “百分之四十”. Are there any specific formatting rules? (Severity: No error)
GPT5: No problem detected
- 8 “But now, aging is fiercest in the high-productivity economies of the Pacific Rim.”
“但是现在人口老龄化在环太平洋地区生产力最高的经济体最为激烈。”
Trainee: “high-productivity” does not mean the highest productivity. (Severity: Minor)
GPT5: style problem: unnatural expression (Severity: Minor)
Trainee-generated error type is more reasonable. It should be a mistranslation instead of an unnatural expression.
Trainee: “老龄化 ... 最为激烈 “ is an unnatural expression (Severity: Minor)
GPT5: word choice nuance (Severity: Minor)
Both error types are acceptable: unnatural expression vs. word choice nuance.
- 9 “Asia has the people to fuel growth, but 90 percent of the expected increase in Asia’s non-farm workforces by 2050 will be where productivity is relatively low.”
“亚洲拥有发展所需的人口，但2050年前亚洲非务农工作力的预期增长的90%将处于生产力相对低的地区。”
Trainee: “non-farm workforces” → “非务农工作力” is a mistranslation (Severity: Minor)
GPT5: bad term choice (Severity: Major)
Both error types are acceptable: unnatural expression vs. word choice nuance.
- 10 “Can Asia shift value chains to where the workers are and raise productivity everywhere?”
“Can Asia shift value chains to where the workers are and raise productivity everywhere?”
Trainee: “everywhere” was mistranslated: should be “地区”, not “方面”. (Severity: Minor)
GPT5: Mis-translation (Severity: Major)
Both error types are acceptable: unnatural expression vs. overly literal phrasing.

	在的地区，并在所有方面都提高生产力呢？”	meaning error.
11	“But now, it could find itself in the crosshairs of trade tensions.” “然而目前亚洲也发现自己在紧张的贸易关系中成了靶子。”	Trainee: Translation for “find itself in” (“发现自己……”) is too literal (Severity: Minor) GPT5: Mis-translation (Severity: Major) Disagreement on error types: “too literal” versus mistranslation (meaning error). “too literal” is more reasonable.
12	“But the value created by tech is moving away from hardware and into new transversal technologies. So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “然而创造价值的科技产业正逐渐从硬体转变为新式横向科技，因此问题在于亚洲是否能在科技创新和科技制造业这些领域进行自我重塑。”	Trainee: The translation of “hardware” (“硬体”) is an expression used in Taiwan. But it is not a conventional expression in mainland China. (Severity: Minor) GPT5: No problem detected.
13	“Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies.” “亚洲有大批的年轻工人，当中有些经济体的生产力又有所提高，这些在很大程度上都是最有利的人口特征。”	Trainee: “surging” means increasing significant. The degree of improvement was not translated properly. (Severity: Major) GPT5: Unnatural/awkward expression (Severity: Minor) Disagreement on error type: The trainee is correct. The issue is a mistranslation rather than an unnatural expression.
14	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” “我们认为全球正处在一个新时代的风口浪尖，各种深层力量相互角逐。而这几股力量与过去30年间塑造世界的力量截然不同。	Trainee: “相互角逐” is added meaning that does not exist in the source sentence. (Severity: Minor) GPT5: meaning addition (Severity: Minor) Same error type
15	“Asia will be at the nexus of these forces and will experience heightened versions of the trends affecting the rest of the world.” “亚洲则将处于这几股力量的角力旋涡中，更深刻地体会这种发展趋势，进而影响着世界其他地方。	Trainee: Mis-translation on “nexus” (center of things) because “旋涡” carries the meaning of struggling. (Severity: Major) GPT5: Mis-translation (Severity: Major) Same error type

		the world" is a mis-translation. (Severity: Major) GPT5: Mis-translation (Severity: Major) Same error type
16	“Asia is the world’s trade crossroads. It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “亚洲是全球贸易的岔道口。这里盘踞着全球20条增长最快的贸易走廊中的18条，20条规模最大的贸易走廊中的13条。”	Trainee: Crossroad-> “岔道口” is a mis-translation (Severity: Major) GPT5: Incorrect term choice (mis-translation) (Severity: Major) Both error types are acceptable: Incorrect term choice leads to mistranslation. Trainee: Need a connection word after the comma. (severity: Minor) GPT5: Grammar/Punctuation (severity: Minor) Equivalent error type
17	“But now, it could find itself in the crosshairs of trade tensions.” “但如今，该地区却被卷入盘根错节中。”	Trainee: Mis-translate “in the crosshairs” (Severity: Major) GPT5: Over-translation (Severity: Major) Equivalent error type: both point to meaning error.
18	“So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “问题就变成亚洲是否能够不仅充当技术生产者，还能将自己重塑成为一个技	Trainee: “reinvent itself” means reinvent for both technology creator and manufacturer. So the verb should cover both. (Severity: Minor) GPT5: No problem detected.
19	“Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies.” “亚洲从很大程度上一直拥有人口红利，在某些经济体中拥有大量的年轻工人和高涨的生产力。”	Trainee: “Demographic sweet spot” translated as “人口红利”, is a mis-translation. GPT5: No problem detected. Trainee: “在某些经济体中拥有” is unnatural Chinese expression. (Severity: Minor) GPT5: Unnatural flow (Severity: Minor) Equivalent error type
20	“But now, aging is fiercest in the high-productivity economies of the Pacific Rim.” “但如今，环太平洋的高效经济体中的老龄化问题也最为严重。”	Trainee: Added meaning with the word “也” (“as well”). (Severity: Minor) GPT5: added meaning (Severity: Minor) Same error type
21	“but 90 percent of the expected increase in Asia’s non-farm workforces by 2050 will be where productivity is relatively low.” “但在截至2050年的亚洲非农劳动力预	Trainee: “90%的增长” (90% increase) is redundant. “90%” is sufficient. (Severity: Minor) GPT5: unnecessary redundancy reduces fluency (Severity: Minor) Same error type

力相对较低的地区。”

- 22 “Asia will be at the nexus of these forces and will experience heightened versions of the trends affecting the rest of the world..”
“亚洲将成为这些力量的纽带，在影响世界其他地区的趋势上，也将经历更高版本的转机。”
- 23 “It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.”
“20个增速最快贸易通道中的18个位于亚洲，20个最大贸易通道中的13个同样起源于亚洲。”
- 24 “But now, it could find itself in the crosshairs of trade tensions.”
“但如今，亚洲发现自己正处于贸易压力的十字准线上。”
- 25 “Asia’s integration has come about through commercial pragmatism rather than political alignment.”
“亚洲的整合来自于商业的实用主义而不是政策导向。”
- 26 “Can that continue if trade tensions rise?”
“如果贸易压力增高，这种整合还会继续吗？。”
- 27 “Asia excels in manufacturing new technology in four areas in particular: consumer electronics, industrial electronics, electric vehicles, and semiconductors”
“亚洲在新科技生产方面遥遥领先，尤其在消费电子产品、工业电子、电动汽车和半导体四个方面。”

Trainee: “Trend”, translated as “转机”, is a mis-translation.
(Severity: Critical)
GPT5: Mis-translation (Severity: Critical)
Same error type

Trainee: “条” is a more appropriate quantity word.
(Severity: Minor)
GPT5: No problem detected.

Trainee : “亚洲” was repeated, which is unnatural. It is better to use a different word “亚洲” (“this place”).
(Severity: Minor)
GPT5: No problem detected.

Trainee: “Tension”, translated as “压力”, is a mis-translation.
(Severity: Major)
GPT5: Mis-translation (Severity: Major)
Same error type
Trainee: “crosshairs”, directly translated as “十字准线”, is awkward in the target language.
(Severity: Minor)
GPT5: Incorrect term choice (Severity: Major)
Both error types are acceptable: incorrect term choice leads to awkwardness in the target language

Trainee: “Political alignment”, translated as “政策导向” is a mis-translation.
(Severity: Major)
GPT5: Mis-translation (Major)
Same error types

Trainee: “trade tension”, translated as “贸易压力”, is a mis-translation.
(Severity: Major)
GPT5: Mis-translation (Major)
Same error types

Human: The source text only mentioned “technology” (“技术”). In contrast, the target text phrase “科技” covers both “science” and “technology”. Therefore, there is added meaning.
(Severity: Minor)
GPT5: No problem detected.

reinvent itself as a tech creator as well as a tech manufacturer.”	clause used the verb “作为”, it would be better to use the same verb in the second clause as well.
“如此一来, 问题就变成亚洲是否能不	(Severity: Minor)
仅作为技术生产者, 也是技术创新者进行再次创新。”	GPT5: “Re-invent itself”→ “进行再次创新” is a major mistranslation. The structure “不仅作为技术生产者, 也是技术创新者进行再次创新” is syntactically awkward.
	(Severity: Major)
	Disagreement on error type: there is a major mistranslation that is identified by GPT5.

Table 1: Negative exemplars (Human trainees versus GPT5). Agreement rate= 28/36 (77.8%). Agreement refers only to whether an error exists, not to its type or severity.

#	Source and Target Sentences	Human & GPT5 generated feedback
1	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” “我们相信新的时代马上就要来临, 这个新时代深层暗涌的力量与在过去30年促成世界格局的力量大不相同。”	Trainee: “the world is on the cusp of a new era” → “促成世界格局”is excellent. GPT5: It is an appropriate terminology choice.
2	“Asia is the world’s trade crossroads. It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “亚洲是世界的贸易交叉衔接地, 坐拥20条发展最快的贸易通道中的18条, 和20条最大的贸易通道中的13条。”	Trainee: “crossroads” → “交叉衔接地”is excellent translation. GPT5: It is a highly accurate translation to convey faithful meaning. Trainee: “It is home to...” → “坐拥” is a good word to use in the context. GPT5: strength in style; slightly elevated and fluent style that avoids rigid literalness while remaining accurate.
3	“So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “因此, 问题在于亚洲是否能向兼顾科技创造和科技制造转型。”	Trainee: The phrase “兼顾” is well selected. GPT5: strength in style; avoids awkward literalism, making the sentence feel native rather than translated.
4	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” “我们相信世界正处于新时代的开端, 与过去30年相比, 影响世界的根本势力已变得非常不同。”	Trainee: “underlying force”→” 根本势力” is a good target language expression. GPT5: a highly accurate terminology. It conveys structural or fundamental drivers rather than superficial changes, aligning well with the analytical tone of the original text.
5	“Asia’s integration has come about through commercial pragmatism rather than political alignment. Can that continue if trade tensions rise?” “亚洲一体化通过商业实用主义而非政治结盟所产生的。一旦贸易紧张局势升级, 一体化的进程还能继续下去吗?”	Trainee: “一体化的进程还能继续下去吗?” is a smooth sentence structure. GPT5: Good at following conventional linguistic convention, and the idiomatic and engaging translation reads smoothly and naturally.
6	“So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “因此问题在于亚洲是否能在科技创新和科技制造业这些领域进行自我重	Trainee: “reinvent itself”→ “自我重塑” is a good target phrase choice. GPT5: accurate translation captures the idea of transformation rather than mere

7	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” The translated target text sentence: “我们认为全球正处在一个新时代的风口浪尖，各种深层力量相互角逐。而这几股力量与过去30年间塑造世界的力量截然不同。”	Trainee: “the world is on the cusp of a new era” → “新时代的风口浪尖” is a vivid translation. GPT5: add rhetorical strength and make the sentence vivid, while still staying close to the original meaning. It does not feel mechanical or overly literal.
8	“Asia is the world’s trade crossroads. It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “亚洲是全球贸易的岔道口。这里盘踞着全球20条增长最快的贸易走廊中的18条，20条规模最大的贸易走廊中的13条。”	Trainee: “home to” → “盘踞” is a good word choice. GPT5: No comment on the word choice.
9	“Asia’s integration has come about through commercial pragmatism rather than political alignment.” “亚洲的整合仰赖的是商业实用主义，而非政治联盟。”	Trainee: “come about through” → “仰赖” is a good word choice. GPT5: The causal nuance of “has come about through... rather than...” is accurately captured by “仰赖的是...而非...”. It is correct and idiomatic in Chinese.
10	“Can that continue if trade tensions rise?” “这条道路还行得通吗?”	Trainee: The target sentence follows the source sentence and is very natural. GPT5: The core meaning of the source sentence is well preserved. The style is the strongest aspect. The phrasing is natural, idiomatic, and fluent in Chinese. It avoids overly technical or literal phrasing and instead uses a familiar metaphor (“道路”) that resonates well in Chinese discourse.
11	“In these places, Asia has more than 40 percent of the revenues, R&D, and patents.” “在这些领域，亚洲坐拥超过40%的利润、研发和专利。”	Trainee: has → “坐拥” is a good word choice. GPT5: it is particularly effective in terms of style: “坐拥” adds a sense of possession and prominence that reads naturally and fluently in Chinese, avoiding a stiff or overly literal rendering. It enhances readability while preserving the emphasis of the original sentence.
12	“Asia has the people to fuel growth, but 90 percent of the expected increase in Asia’s non-farm workforces by 2050 will be where productivity is relatively low.” “亚洲虽然拥有人力来推动增长，但在截至2050年的亚洲非农劳动力预期增长中，其中90%的增长都来自生产力相对较低的地区。”	Trainee: “其中90%的增长” is a good translation. GPT 5.4: the style is strong: creates a smooth, idiomatic flow rather than a rigid, word-for-word rendering. It reads naturally and fluently in Chinese
13	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” “人们认为相较于过去30年塑形世界的力量，新时代截然不同的潜在力量，将世界推向风口浪尖。”	Trainee: “on the cusp of” → “推向风口浪尖” is a vivid translation. GPT5: strong style (fluency and naturalness): effectively convey a sense of urgency and transformation, making the sentence engaging and vivid rather than stiff or overly literal.

	and will experience heightened versions of the trends affecting the rest of the world.” “亚洲将成为这些力量的纽带，在影响世界其他地区的趋势上，也将经历更高版本的转机。”	GPT5: the term “纽带” is an appropriate and understandable choice for a general audience when translating “nexus,” avoiding overly technical or obscure wording.
15	“Asia is the world’s trade crossroads. It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “亚洲是世界贸易枢纽。20个增速最快贸易通道中的18个位于亚洲，20个最大贸易通道中的13个同样起源于亚洲。”	Trainee: “trade crossroads”→“贸易枢纽” is an excellent target phrase choice. GPT5: Key economic and trade-related terms are well chosen. “贸易枢纽” appropriately reflects “trade crossroads,” aligning with standard usage in economic and policy contexts.
16	“Asia excels in manufacturing new technology in four areas in particular: consumer electronics, industrial electronics, electric vehicles, and semiconductors.” “亚洲在新科技生产方面遥遥领先，尤其在消费电子产品、工业电子、电动汽车和半导体四个方面。”	Trainee: “excel”→“遥遥领先” is an excellent choice. GPT5: No comment on the word choice.
17	“But the value created by tech is moving away from hardware and into new transversal technologies.” “然而技术产生的价值正从硬件生产转移到新兴交叉科技。”	Trainee: “transversal technologies”→“交叉科技” is an excellent target phrase choice. GPT5: The rendering of “transversal technologies” as “交叉科技” is particularly strong. It reflects an understanding of the concept as cross-domain or interdisciplinary technologies, which is consistent with how such ideas are expressed in Chinese technical and business contexts.
18	“But now, aging is fiercest in the high-productivity economies of the Pacific Rim.” “可是如今，老龄化在太平洋地区高生产力的经济体中成为最棘手的问题。”	Trainee: “fiercest”→“最棘手的问题” is an very interesting translation. GPT5: The core meaning of the source sentence is well preserved. “aging is fiercest” is effectively rendered as “老龄化...成为最棘手的问题,” which captures the intensity and seriousness of the demographic challenge, even though it is slightly interpretive. The relationship between aging and “high-productivity economies of the Pacific Rim” is clearly and correctly conveyed.

Table 2: Positive exemplars (Human trainees versus GPT5). Agreement rate 16/18 (88.8%). Agreement refers only to whether a positive translation exists. Human trainees’ comments tend to be shorter and judgment-based while GPT5 comments are more analytical and evaluative.

#	Source and Target Sentences	GPT5 generated feedback (highlighted are false negatives identified by humans)
1	“We believe that <i>the world</i> is on the cusp of a new era...” “我们相信新的时代马上就要来临...”	Accuracy/mistranslation issue: “on the cusp of” →“马上就要来临” [Human evaluators’ disagreement: The translation is acceptable as long as it reflects the sentiment of “change is about to happen” or “a new era is about to start”.] Omission: “the world” is omitted. Style: slightly informal; less precise tone.

		Mistranslation: “in the crosshairs of trade tensions”→ “得到了格外强烈的体现”. Omission: “could find itself” Addition: “也” [Human evaluators’ disagreement: We don’t think it’s an addition here. 也 isn’t “also” all the time. Here it looks like 也 is used to continue on the context of the previous sentence (Asia is centre of trade paths, it may now be the centre of trade tensions).]
2	“But now, it could find itself in the crosshairs of trade tensions.” “当今贸易 紧张局势 也在这里得到了格外强烈的 体现”	Style: Unnatural / awkward expression Awkward referential expression (minor): “这里” is vague and weak.
3	“Asia’s integration has come about through commercial pragmatism rather than political alignment.” “亚洲的融入是通过商业务实性，而不是政治 取向 一致性。”	Mistranslation: integration”→ “融入” Terminology error: “political alignment”→ “政治取向一致性” Style – Unnatural / awkward expression: : “是通过商业务实性，而不是政治取向一致性” is very literal and stiff. Linguistic convention issue: “通过商业务实性” is grammatically acceptable but unnatural. [Human evaluators’ disagreement: 基于/源于 are better but 通过 isn’t wrong and still flies, We wouldn’t flag this as an error.]
4	“Can that continue if trade tensions rise?” “亚洲还可以维持现状吗？”	Omission: “if trade tensions rise”. The LLM did not go back and check the context. [Human evaluators’ disagreement: There is no omission here. The translation of “Can that continue if trade tensions rise?” is broken down into 2 clauses in Chinese separated by a comma, the translation is “如果贸易紧张局势加剧，亚洲还可以维持现状吗？” . The alleged omission is placed before the comma, the LLM simply missed it. Translation is not simple mapping, sometimes one sentence in the source is broken down into several in the target, sometimes several clauses can be joined together to form one wholesome sentence in the translation.] Mistranslation: “that continue” is rendered as “维持现状” Style (Slightly awkward phrasing): “that continue” → “维持现状” [Human evaluators’ disagreement: the mistranslation seems to be caused by a misunderstanding of the source sentence and it is not the word “continue”.]
5	“Asia excels in manufacturing new technology in four areas in particular” “亚洲在如下四项新型科技 制造方面超群拔类”	Accuracy issue: “如下四项” (the following four items) Terminology issue: “manufacturing new technology”→ “新型科技制造” Style issue: “如下四项” is slightly formal/awkward, not natural. [Human evaluators’ disagreement: 以下 sounds

		that it should not be counted as an error.] Style issue: “超群拔类” is overly literary and formal, not consistent with the analytical tone of the source text.
6	“consumer electronics, industrial electronics, electric vehicles, and semiconductors.” “消费电子产品、工业电子产品、电动汽车、半导体。”	GPT5 did not find any problem.
7	“Asia has more than 40 percent of the revenues, R&D, and patents.” “亚洲占了超过百分之四十的收益、研发以及专利。”	Terminology issue: “revenues” → “收益” Style issue: “占了超过百分之四十的...” is slightly less natural.
8	“But now, aging is fiercest in the high-productivity economies of the Pacific Rim.” “但是现在人口老龄化在环太平洋地区生产力最高的经济体最为激烈。”	Linguistic convention issue: Missing structural particles (“的”“中”) in “在环太平洋地区生产力最高的经济体最为激烈”. Style issue: “生产力最高的经济体” is an unnatural expression that is overly literal. Style (word choice nuance): “最为激烈” used for “aging”
9	“Asia has the people to fuel growth, but 90 percent of the expected increase in Asia’s non-farm workforces by 2050 will be where productivity is relatively low.” “亚洲拥有发展所需的人口，但2050年前亚洲非务农工作力的预期增长的90%将处于生产力相对低的地区。”	Accuracy issue: “will be where productivity is relatively low” → “将处于生产力相对低的地区” Terminology issue: “non-farm workforces” → “非务农工作力” [Human evaluators’ disagreement: 非务农 and 非农业 both should work. It is not an error]. Linguistic convention issue: “预期增长的90%将处于...” structure is grammatically acceptable but slightly clunky due to repeated “的” and long modifier stacking. That is awkward in Chinese. Style issue: “发展所需的人口” is slightly stiff.
10	“Can Asia shift value chains to where the workers are and raise productivity everywhere?” “亚洲是否可以将价值链转移到工人所在的地区，并在所有方面都提高生产力呢？”	Accuracy issue: “raise productivity everywhere” → “在所有方面都提高生产力” Style issue: “是否可以将价值链转移到...” is slightly literal phrasing
11	“But now, it could find itself in the crosshairs of trade tensions.” “然而目前亚洲也发现自己紧张的贸易关系中成了靶子。”	Accuracy issue: “trade tensions” → “紧张的贸易关系”. The source refers specifically to trade tensions/conflicts, which implies geopolitical friction or disputes. The translation weakens the meaning and changes the nuance. [Human evaluators’ disagreement: The turn of phrase is a bit awkward but it for Machine Translation in the Americas,

		over translation for the exam.] Accuracy issue: “find itself in the crosshairs” is rendered as “发现自己...成了靶子”. Omission/meaning shift: The modal “could” (indicating possibility) is not clearly conveyed. Grammar: The phrase “紧张的贸易关系中成了靶子” is grammatically awkward. And there are two related style problems.
12	“But the value created by tech is moving away from hardware and into new transversal technologies. So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “然而创造价值的科技产业正逐渐从硬体转变为新式横向科技，因此问题在于亚洲是否能在科技创新和科技制造业这些领域进行自我重塑。”	Accuracy issue: changes the subject from “value created” to “the tech industry”. Incorrect term: “横向科技” is not a standard Chinese rendering. Accuracy issue: “tech creator”→ “科技创新” Entire sentence structure is an over-expansion not present in source. [Human evaluators’ disagreement: The source words “creator” and “manufacturer” are stylistic choice. The “base” meaning of the segment is “create/invent tech” and “manufacture tech”, which is preserved. We would not expect the parts of speech in the translation to align perfectly with the source - syntax works differently in different languages.]
13	“Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies.” “亚洲有大批的年轻工人，当中有些经济体的生产力又有所提高，这些在很大程度上都是最有利的人口特征。”	Unnatural collocation: “新式横向科技”. Rigid translation: “因此问题在于” Accuracy issue: “demographic sweet spot” → “最有利的人口特征”. Accuracy issue: 这些在很大程度上都是” (these largely are) introduces an explanatory structure not present in the source sentence. [Human evaluators’ disagreement: We think 很大程度上 matches “largely” in the source text, what’s changed is the sentence structure - but the meaning of “largely” is preserved. This is not an error.] Unnatural flow: “这些在很大程度上都是最有利的人口特征” is awkward and not idiomatic in Chinese.
14	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” “我们认为全球正处在一个新时代的风口浪尖，各种深层力量相互角逐。而这几股力量与过去30年间塑造世界的力量截然不同。”	Over-translation/Addition: “On the cusp”→“风口浪尖” Style: “新时代的风口浪尖” is rhetorically strong and journalistic; unnatural or inappropriate metaphorical tone Linguistic conventions: The translation splits the original single sentence into two sentences
15	“Asia will be at the nexus of these forces and will in turn influence other places.”	Accuracy issue: “进而影响着世界其他地方” means “and in turn influence other places”

	the trends affecting the rest of the world.” “亚洲则将处于这几股力量的角力旋涡中，更深刻地体会这种发展趋势，进而影响着世界其他地方。”	Accuracy issue: “nexus”→ “角力旋涡” Style: Unnatural / Overly Dramatic Expression. “角力旋涡” and “更深刻地体会” Awkward expression: 更深刻地体会这种发展趋势”
16	“Asia is the world’s trade crossroads. It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “亚洲是全球贸易的岔道口。这里盘踞着全球20条增长最快的贸易走廊中的18条，20条规模最大的贸易走廊中的13条。”	Accuracy issue: “home to”→“盘踞着” Incorrect term choice: “crossroads”->“岔道口” Style: the second sentence is an unnatural expression. Linguistic conventions: lack parallel structure clarity.
17	“But now, it could find itself in the crosshairs of trade tensions.” “但如今，该地区却被卷入盘根错节中。”	Accuracy issue: The phrase “盘根错节的” (complicated) introduces an additional semantic element that does not exist in the source text. Style: The phrase “盘根错节的贸易紧张局势” sounds stylistically heavy and somewhat literary compared with the neutral analytical tone of the source text.
18	“So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “问题就变成亚洲是否能够不仅充当技术生产者，还能将自己重塑成为一个技术创新者。”	Accuracy issue: “So, the question is”→ “问题就变成” Incorrect term choice: “tech manufacture”→ “技术生产者” Style: The sentence structure is somewhat heavy and verbose compared with standard Chinese economic writing.
19	“Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies.” “亚洲从很大程度上一直拥有人口红利，在某些经济体中拥有大量的年轻工人和高涨的生产力。”	Accuracy issue: “surging productivity”→“高涨的生产力” Incorrect term choice: “productivity”→“生产力”: “productivity” is translated as “生产力” instead of the more precise economic term “生产率”. [Human evaluators’ disagreement: We don’t think 生产力 is wrong. Here the paragraph is talking about demographics and labour, and with young workers it naturally factors that the output capability is high. “Productivity” in this context refers more likely to the ability to produce, rather than the ratio of material output over input.] Style: “从很大程度上一直拥有” is somewhat awkward and verbose. Style: “大量的年轻工人” is understandable but slightly literal.
20	“But now, aging is fiercest in the high-productivity economies of the Pacific Rim.”	Accuracy issue: The word “也” (also) is added in the translation, but the source sentence does not contain the meaning “also.” [Human evaluators’ disagreement: We

	老龄化问题也最为严重。”	one is to back-translate this into English. However, there is an implied contrast between this sentence and the previous sentence in the source. “Asia population situation good. (But) some countries have population problems”. 也 is used as an emphasis for the aging issue, which is the point of contrast. Nothing is wrong]. Incorrect terminology: “high productivity economics”→ “高效经济体” Style: The structure “环太平洋的高效经济体中的老龄化问题” is somewhat heavy and indirect compared with natural Chinese. Accuracy issue: “will be where productivity is relatively low” → “来自生产力相对较低的地区”. The source sentence means the workforce increase will occur in low-productivity areas, not that it comes from those areas. “来自” introduces an unintended source meaning.
21	“but 90 percent of the expected increase in Asia’s non-farm workforces by 2050 will be where productivity is relatively low.” “但在截至2050年的亚洲非农劳动力预期增长中，其中90%的增长都来自生产力相对较低的地区。”	[Human evaluators’ disagreement: We think this is fine and would not cause misunderstanding on the part of the reader. The core message refers to where labor growth is, the reader would think about location rather than where the people are born (“from”).] Accuracy issue: “by 2050”→ “截至2050年” (slightly alters the meaning) Style: Repetition in “预期增长中，其中90%的增长” creates redundancy and slightly reduces fluency. Style: “亚洲非农劳动力预期增长中” is awkward phrasing
22	“Asia will be at the nexus of these forces and will experience heightened versions of the trends affecting the rest of the world..” “亚洲将成为这些力量的纽带，在影响世界其他地区的趋势上，也将经历更高版本的转机。”	Incorrect term: “Nexus” → “纽带” Structure: “在影响世界其他地区的趋势上” is grammatically awkward. Accuracy issue: “heightened versions”→ “更高版本”. Accuracy issue: “trends”→ “转机” Style: Overall phrasing is unnatural and overly literal.
23	“It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “20个增速最快贸易通道中的18个位于亚洲，20个最大贸易通道中的13个同样起源于亚洲。”	Inaccuracy: “home to”→ “起源于亚洲” Incorrect term choice: “trade corridors”→ “贸易通道” Style: Use of “同样” is awkward/unnecessary emphasis
24	“But now, it could find itself in the crosshairs of trade tensions.” “但如今，亚洲发现自己正处于贸易压力的十字准线上。”	Inaccuracy: “trade tensions” → “贸易压力” Incorrect term choice: “crosshairs” → “十字准线” Style: “find itself”→ “发现自己正处于” Style: “贸易压力的十字准线上” is stylistically unnatural and reads like a word-for-word rendering rather than

25	“Asia’s integration has come about through commercial pragmatism rather than political alignment.” “亚洲的整合来自于商业的实用主义而不是政策导向。”	Accuracy issue: “political alignment” is translated as “政策导向” Style: “来自于商业的实用主义” is slightly stiff and literal Suboptimal term choice: “商业的实用主义”
26	“Can that continue if trade tensions rise?” “如果贸易压力增高，这种整合还会继续吗？。”	Accuracy issue: “trade tensions” → “贸易压力” Style: “这种整合还会继续吗” is understandable but slightly stiff and literal Accuracy issue: “excels in” → “遥遥领先” Style: “新科技生产” is not a natural Chinese collocation.
27	“Asia excels in manufacturing new technology in four areas in particular: consumer electronics, industrial electronics, electric vehicles, and semiconductors” “亚洲在新科技生产方面遥遥领先，尤其在消费电子产品、工业电子、电动汽车和半导体四个方面。”	Accuracy issue: “consumer electronics” → “消费电子产品” Adding “产品” slightly expands the meaning. While common in Chinese, it is not strictly present in the source text. [Human evaluators’ disagreement: Such term exists and is used in Chinese. It is also the term of choice in 3 of the 4 translations (the other one being 消费者电子产品), but the MQM did not pick out the term until version 4.] Accuracy issue: “reinvent itself” → “进行再次创新” Addition: The phrase “如此一来” (“thus/therefore”) is added but does not appear in the source sentence. [Human evaluators’ disagreement: We don’t think the translation is wrong. It is likely serves the function “so” serves in the source. “Given the above context” is the idea “so” and 如此一来 express. The translation is appropriate.]
28	“So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “如此一来，问题就变成亚洲是否能不仅作为技术生产者，也是技术创新者进行再次创新。”	Style: The structure “不仅作为技术生产者，也是技术创新者进行再次创新” is syntactically awkward. Style: The phrase “问题就变成” is slightly heavy and literal. Terminology: “tech manufacturer” → “技术生产者”

Table 3: True and false negative exemplars generated by GPT5. Percentage of false negatives found by human evaluators is 15/92 (16.3%).

#	Source and Target Sentence	GPT5 generated feedback (highlighted are false positives identified by humans)
1	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” “我们相信新的时代马上就要来临，这个新时代深层暗涌的力量与在过去30年促成世界格局的力量大不相同。 “Asia is the world’s trade crossroads. It’s	“that shaped our world” → “促成世界格局”: The choice of “世界格局” for “our world” is appropriate in this context, as it reflects a macro-level, analytical perspective consistent with the tone of the source.

	corridors and 13 of the 20 largest.” “亚洲是世界的贸易交叉衔接地，坐拥20条发展最快的贸易通道中的18条，和20条最大的贸易通道中的13条。”	meaning is conveyed very faithfully. “trade crossroads” is effectively rendered as “贸易交叉衔接地,” capturing the idea of Asia as a central hub.
3	“So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “因此，问题在于亚洲是否能向兼顾科技创造和科技制造转型。”	“as well as”→“兼顾：The phrase “向兼顾...转型” is concise and avoids awkward literalism, making the sentence feel native rather than translated. “underlying force” →“根本势力”: The phrase “very different underlying forces” is effectively translated as “根本势力已变得非常不同,” preserving the conceptual emphasis on foundational drivers of change. There is no omission or distortion of meaning. [Human evaluators’ disagreement: Translating “underlying” as 根本 is not wrong enough to warrant a point deduction, but 根本 can mean “source/origin” whereas the meaning of “underlying” is more like “deep down/deep within”. Therefore, this is a minor false positive.]
4	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” “我们相信世界正处于新时代的开端，与过去30年相比，影响世界的根本势力已变得非常不同。”	
5	“Asia’s integration has come about through commercial pragmatism rather than political alignment. Can that continue if trade tensions rise?” “亚洲一体化通过商业实用主义而非政治结盟所产生的。一旦贸易紧张局势升级，一体化的进程还能继续下去吗？”	“Can that (integration) continue if trade tensions rise”→“一体化的进程还能继续下去吗？”: Linguistic Conventions The sentence is grammatically sound, with correct punctuation and no typographical issues. The use of the question mark and the conditional structure (“一旦...还能...吗”) is properly handled. Style The translation reads smoothly and naturally in Chinese, particularly the second sentence, which is idiomatic and engaging. The rhetorical question structure is well rendered and aligns with native Chinese discourse patterns.
6	“So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “因此问题在于亚洲是否能在科技创新和科技制造业这些领域进行自我重塑。”	“reinvent itself”→“自我重塑”: the notion of “reinvent itself” is effectively rendered as “进行自我重塑,” which captures the idea of transformation rather than mere adjustment. [Human evaluators’ Disagreement: 重塑 is used well, but 自我 personifies the subject Asia too much (not intended in source text).]
7	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” The translated target text sentence: “我们认为全球正处在一个新时代的风口浪尖，各种深层力量相互角逐。而这几股力量与过去30年间塑造世界的力量截然不同。”	“the cusp of a new era”→“新时代的风口浪尖”: The core meaning of the source sentence is conveyed clearly and faithfully. The notion of “on the cusp of a new era” is effectively captured by “正处在一个新时代的风口浪尖,” which reflects a moment of transition and significance. [Human evaluators’ disagreement: Net negative. The idiom is good (so the comment about it being lively isn’t wrong). But it does not accurately

threshold” or “change is about to happen”, the segment about the underlying forces describes the subject of “on the cusp”. 风口浪尖 is more on the side or “at the height of tensions”; it signifies underlying struggle, which is not suggested in the source text and therefore distorts the meaning.

Translation can be subjective. There are readers who marvel at the use of fancy words without thinking deeper at the meaning. 用词漂亮差不多得了. But we think here since the meaning is not fully reflective, the negative outweighs the positive.]

8 “Asia’s integration has come about through commercial pragmatism rather than political alignment.”
“亚洲的整合仰赖的是商业实用主义，而非政治联盟。”

“come about”→“仰赖”: the causal nuance of “has come about through... rather than...” is well captured by “仰赖的是...而非...”.

9 “Can that continue if trade tensions rise?”
“这条道路还行得通吗?”

“Can that continue...”→ “这条道路还行得通吗?”:

Accuracy: The core meaning of the source sentence is well preserved. The rhetorical question “Can that continue...” is effectively conveyed through “还行得通吗,” which captures the sense of feasibility and continuity under changing conditions. While “that” is slightly interpreted as “这条道路,” this is a justified and contextually appropriate explication rather than an error.

Style: This is the strongest aspect. The phrasing is natural, idiomatic, and fluent in Chinese. “这条道路还行得通吗” reads like original Chinese writing rather than a translation, and it maintains the rhetorical tone of the source text effectively.

Audience Appropriateness: The sentence is clear and accessible to educated general readers. It avoids overly technical or literal phrasing and instead uses a familiar metaphor (“道路”) that resonates well in Chinese discourse.

10 “In these places, Asia has more than 40 percent of the revenues, R&D, and patents.”
“在这些领域，亚洲坐拥超过40%的利润、研发和专利。”

“has”→“坐拥”:

Style: The phrase “亚洲坐拥超过40%” is particularly effective. “坐拥” adds a sense of possession and prominence that reads naturally and fluently in Chinese, avoiding a stiff or overly literal rendering. It enhances readability while preserving the emphasis of the original sentence.

11 “Asia has the people to fuel growth, but 90 percent of the expected increase in Asia’s non-farm workforces by 2050 will be where productivity is relatively low.”
Proceedings of the 17th Conference of the Association

“90 percent of the expected increase”→“其中90%的增长”:

The translation reads naturally and fluently in Chinese. The use of “截至”

	截至2050年的亚洲非农劳动力预期增长中，其中90%的增长都来自生产力相对较低的地区。”	自…” creates a smooth, idiomatic flow rather than a rigid, word-for-word rendering. The sentence maintains a formal, analytical tone consistent with the original text. [Human evaluators’ disagreement: Neutral, the translation is correct, but nothing outstanding, and using 增长 twice in close succession may seem redundant to some.]
12	“We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” “人们认为相较于过去30年塑形世界的力量，新时代截然不同的潜在力量，将世界推向风口浪尖。”	“the cusp of a new era “→“新时代的 风口浪尖”: The sentence reads smoothly and idiomatically in Chinese. Expressions like “风口浪尖” effectively convey a sense of urgency and transformation, making the sentence engaging and vivid rather than stiff or overly literal. [Human evaluators’ disagreement: Translation can be subjective. There are readers who marvel at the use of fancy words without thinking deeper at the meaning. But We think here since the meaning is not fully reflective, the negative outweighs the positive.]
13	“Asia will be at the nexus of these forces and will experience heightened versions of the trends affecting the rest of the world.” “亚洲将成为这些力量的纽带，在影响世界其他地区的趋势上，也将经历更高版本的转机。”	“nexus”→”纽带”: The term “纽带” is an appropriate and understandable choice for a general audience when translating “nexus,” avoiding overly technical or obscure wording.
14	“Asia is the world’s trade crossroads. It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “亚洲是世界贸易枢纽。20个增速最快贸易通道中的18个位于亚洲，20个最大贸易通道中的13个同样起源于亚洲。”	“crossroad”→“枢纽”: Key economic and trade-related terms are well chosen. “贸易枢纽” appropriately reflects “trade crossroads,” and “贸易通道” is a suitable and consistent rendering of “trade corridors,” aligning with standard usage in economic and policy contexts.
15	“But the value created by tech is moving away from hardware and into new transversal technologies.” “然而技术产生的价值正从硬件生产转移到新兴交叉科技。”	“transversal technologies”→“交叉科技”: The rendering of “transversal technologies” as “交叉科技” is particularly strong. It reflects an understanding of the concept as cross-domain or interdisciplinary technologies, which is consistent with how such ideas are expressed in Chinese technical and business contexts. [Human evaluators’ disagreement: this term does not seem adopted in Chinese and does not make much sense at face value.]
16	“But now, aging is fiercest in the high-productivity economies of the Pacific Rim.” “可是如今，老龄化在太平洋地区高生产力的经济体中成为最棘手的问题。”	“fiercest”→“最棘手的问题”: The core meaning of the source sentence is well preserved. “aging is fiercest” is effectively rendered as “老龄化…成为最棘手的问题,” which captures the intensity and seriousness of the demographic challenge, even though it is slightly interpretive. The relationship

clearly and correctly conveyed.
[Human evaluators' disagreement: The translation is idiomatic, but it also distorts the meaning.]

Table 4: True and false positive exemplars generated by GPT5. Percentage of false positive found by human evaluators is 7/19 (36.84%)

#	Different opinions from human trainees (comment 1) and GPT5.4 (comment 2)	DeepSeek arbitrations and human evaluator's final judgments
1	There are two comments on the translation pair below: English: "consumer electronics, industrial elec-tronics, electric vehicles, and semiconductors." Chinese: "消费电子产品、工业电子产品、电动汽车、半导体。" Comment 1: The last item in a list must be connected with a "和". Comment 2: No problem detected. Which comment is better? Why?	DeepSeek: Comment 1 is better. The English list explicitly includes "and"; before the last item, so the Chinese translation should reflect that with a conjunction like "和"; for naturalness and fidelity. While omitting it is not strictly incorrect in all contexts, the "must" in Comment 1 is slightly overstated, but it correctly identifies the improvement needed. Comment 2 overlooks this common convention. DeepSeek agrees with the human trainee. Human evaluators: disagree with DeepSeek: Such a convention does not exist. DeepSeek is wrong!
2	There are two comments on the translation pair below: "Asia has more than 40 percent of the revenues" "亚洲占了超过百分之四十的收益" Comment 1: Why "40%" " should be written in Chinese text "百分之四十". Are there any specific formatting rules? Comment 2: No problem detected. Which comment is better? Why?	DeepSeek: Comment 2 is better because it correctly identifies that the translation is accurate and follows standard Chinese localization practices. In Chinese, percentages are commonly written as "百分之四十"; rather than "40%"; in formal or narrative text, especially when the number is written out in full. Comment 1's concern is unnecessary, as there is no specific formatting rule violated—this usage is perfectly natural and idiomatic in Chinese. DeepSeek agrees with GPT5.4: this usage is perfectly natural and idiomatic in Chinese. Human evaluators: agree with DeepSeek. DeepSeek is right.
3	There are two comments on the translation pair below: "But the value created by tech is moving away from hardware and into new transversal technologies. So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer." "然而创造价值的科技产业正逐渐从硬体转变为新式横向科技, 因此问题在于亚洲是否能在科技创新和科技制造业这些领域进	DeepSeek: Comment 1 is better because it accurately identifies a regional terminology issue. The word "硬件" is standard in mainland China, while "硬体" is primarily used in Taiwan. Since the translation likely targets a mainland Chinese audience, using "硬体" could be considered unnatural or a regional mismatch. Comment 2 overlooks this nuance.

	Comment 1: The translation of “hardware” (硬件) is an expression used in Taiwan. But it is not a conventional expression in mainland China. Comment 2: No problem detected. Which comment is better? Why?	feedback for improving the translation’s appropriateness. DeepSeek agrees with human trainee’s point: the regional terminology issue exists. Human evaluators: agree with DeepSeek. DeepSeek is right.
4	There are two comments on the translation pair below: “So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” “问题就变成亚洲是否能够不仅充当技术生产者，还能将自己重塑成为一个技术创新者。” Comment 1: “reinvent itself” means reinvent for both technology creator and manufacturer. So the verb should cover both. Comment 2: No problem detected. Which comment is better? Why?	DeepSeek: Comment 2 is better because the translation accurately captures the meaning of the original. In the English sentence, “reinvent itself”; refers specifically to the transformation into a tech creator, given that Asia is already a tech manufacturer. The Chinese translation reflects this by using “充当” (act as) for the existing role and “将自己重塑成为”; (reinvent itself as) for the new role, which correctly separates the two concepts. Comment 1 incorrectly suggests that the verb must cover both roles, but that would be redundant or misleading since the manufacturing role does not require reinvention. DeepSeek agrees with GPT5.4: , “reinvent itself”; refers specifically to the transformation into a tech creator, given that Asia is already a tech manufacturer. Human evaluators: agree with DeepSeek. DeepSeek is right.
5	There are two comments on the translation pair below: “Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies.” “亚洲从很大程度上一直拥有人口红利，在某些经济体中拥有大量的年轻工人和高涨的生产力。” Comment 1: “Demographic sweet spot”. translated as “人口红利”, is a mis-translation. Comment 2: No problem detected. Which comment is better? Why?	DeepSeek: Comment 2 is better because the translation “人口红利”; accurately captures the meaning of “demographic sweet spot”; In economic and demographic contexts, “demographic sweet spot”; refers to a period when a large working-age population relative to dependents creates favorable conditions for economic growth—exactly what “人口红利”; (demographic dividend) denotes. The Chinese version also correctly conveys the additional details about young workers and surging productivity. Therefore, the translation is correct, and Comment 2 is valid. DeepSeek agrees with GPT5.4: “人口红利” is a good translation. Human evaluator: disagrees with DeepSeek: It changes the meaning too much. DeepSeek is wrong!
6	There are two comments on the translation pair below: “It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “20个增速最快贸易通道中的18个位于亚洲，20个最大贸易通道中的13个同样起源于亚洲。” Comment 1: “条” is a more appropriate quantity word.	DeepSeek: Comment 1 is better because the Chinese translation uses the generic measure word “个” (gè) for “贸易通道” (trade corridors), but the more precise and natural measure word for “通道” (corridor) is “条” (tiáo). While “个” is acceptable in informal contexts, “条” is more accurate and idiomatic in

Which comment is better? Why?

describing physical or conceptual routes. Comment 2 overlooks this subtle but important nuance. DeepSeek agrees with human trainee. **Human evaluators:** agree with DeepSeek: “条” is better. DeepSeek is correct.

7 There are two comments on the translation pair below: “It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “20个增速最快贸易通道中的18个位于亚洲，20个最大贸易通道中的13个同样起源于亚洲。” Comment 1: “亚洲” was repeated, which is unnatural. It is better to use a different word “此地” (“this place”). Comment 2: No problem detected. Which comment is better? Why?

DeepSeek: Comment 2 is better. The repetition of “亚洲” in the Chinese translation is not unnatural—it maintains clarity and emphasis in a parallel structure. Using “此地” (“this place”) could introduce ambiguity if the referent (Asia) is not explicitly stated nearby, and it sounds slightly archaic or formal in this context. The original translation is clear and idiomatic, so no change is needed. DeepSeek agrees with GPT5.4. That is not a problem to repeat “亚洲” . **Human evaluators:** agree with DeepSeek. DeepSeek is correct.

8 There are two comments on the translation pair below: ““Asia excels in manufacturing new technology in four areas in particular: consumer electronics, industrial electronics, electric vehicles, and semiconductors” “亚洲在新科技生产方面遥遥领先，尤其在消费电子产品、工业电子、电动汽车和半导体四个方面。” Comment 1: The source text only mentioned “technology” (“技术”). In contrast, the target text phrase “科技” covers both “science” and “technology”. Therefore, there is added meaning. Comment 2: No problem detected. Which comment is better? Why?

DeepSeek: Comment 2 is better because the translation is accurate and natural in Chinese. The term “科技” in Chinese commonly serves as a shorthand for “技术” (technology) in contexts like manufacturing and industry, especially when paired with “新” (new). While “科技” literally covers both science and technology, its usage here is idiomatic and does not introduce unwarranted meaning—the source’s “new technology” is faithfully rendered. Comment 1 is overly pedantic and fails to account for standard Chinese lexical conventions. DeepSeek agrees with GPT5.4: It is a standard Chinese lexical convention. **Human evaluators:** agree with DeepSeek. DeepSeek is correct.

Table 5: DeepSeek arbitrated disagreements on negative exemplars between human trainees and GPT-5.4. Based on the human evaluators' final judgments, DeepSeek's arbitration was correct in 6 of the 8 cases.

#	Different opinions from human trainees (comment 1) and GPT5.4 (comment 2)	DeepSeek’s arbitration and Human evaluator’s Re-evaluation.
1	There are two comments on the translation pair below: “Asia is the world’s trade crossroads. It’s home to 18 of the 20 fastest-growing trade corridors and 13 of the 20 largest.” “亚洲是全球贸易的岔道口。这里盘踞着全球20条增长最快的贸易走廊中的18条，20条规模最大的贸易走廊中的13条。”	DeepSeek: Comment 1 is better because it provides specific, constructive feedback by praising a particular word choice “盘踞”) that effectively captures the meaning and tone of “home”; in the context. It acknowledges a successful translation decision, which is helpful for the translator and encourages good

	word choice. Comment 2: No positive comments. Which comment is better? Why?	Comment 2 offers no insights or justification, making it unhelpful for improving translation quality or understanding. DeepSeek agrees with human trainees that particular word choice is a successful translation decision. Human evaluators: agree with DeepSeek and human trainee. positive, good language quality. 盘踞 is a verb suitable to describe something long and thin, which is in accordance with the concept of “corridor”. DeepSeek is correct.
2	There are two comments on the translation pair below: “Asia excels in manufacturing new technology in four areas in particular: consumer electronics, industrial electronics, electric vehicles, and semiconductors.” “亚洲在新科技生产方面遥遥领先，尤其在消费电子产品、工业电子、电动汽车和半导体四个方面。” Comment 1: “excel”→“遥遥领先” is an excellent choice. Comment 2: No positive comments. Which comment is better? Why?	DeepSeek: Comment 1 is better because it provides specific, constructive feedback that highlights a successful translation choice. The phrase :遥遥领先” effectively captures the meaning of “excels”; in this context, adding emphasis that fits naturally in Chinese. This type of positive comment encourages the translator and offers a clear example of good practice. Comment 2, on the other hand, offers no useful information—it neither acknowledges the quality of the translation nor helps the translator improve. In a review or learning setting, specific praise (or critique) is far more valuable than a generic dismissal. DeepSeek agrees with human trainee’s comment. Human evaluators: agree with DeepSeek: Positive, good language quality. DeepSeek is correct.

Table 6: DeepSeek arbitrated disagreements on positive exemplars between human trainees and GPT-5.4. Based on the human evaluators' final judgments, DeepSeek's arbitration was correct in two of the two cases.

#	GPT5.4 generated explanation (comment 1) and human evaluator’s disagreements (comment 2)	DeepSeek’s arbitration and Human evaluator’s Re-evaluation
1	There are two comments on the translation pair below: English: “We believe that the world is on the cusp of a new era...” Chinese: ““我们相信新的时代马上就要来临...”” Comment 1: Accuracy/mistranslation issue: “on the cusp of” □ “马上就要来临” Comment 2: The translation is acceptable as long as it reflects the sentiment of “change is about to happen” or “a new era is about to start”. Which comment is better? Why?	Deep Seek: Comment 2 is better because it recognizes that the translation effectively conveys the core meaning and sentiment of the original phrase, even if it doesn’t use a literal equivalent of “on the cusp of”; In translation, capturing the intended meaning and naturalness in the target language is often more important than word-for-word accuracy. The Chinese expression “马上就要来临” naturally communicates the immediacy and impending change, which aligns with the original’s sense of being on the verge of a new era. Comment 1 is overly literal and fails to account for

		agrees with human evaluator's judgment. Evaluators' further investigation confirms the original judgment. DeepSeek is correct.
2	There are two comments on the translation pair below: English: "But now, it could find itself in the crosshairs of trade tensions." Chinese: "当今贸易紧张局势也在这里得到了格外强烈的体现" Comment 1: Add meaning: "也" added meaning in Chinese Comment 2: We don't think it's an addition here. 也 isn't "also" all the time. Here it looks like 也 is used to continue on the context of the previous sentence (Asia is centre of trade paths, it may now be the centre of trade tensions). Which comment is better? Why?	DeepSeek: Comment 2 is better because it recognizes the pragmatic and cohesive function of "也" in Chinese discourse. In this context, "也" does not primarily mean "also" in a literal additive sense; rather, it serves as a connective device that ties the current statement to the preceding context (e.g., Asia being the center of trade paths). This is a natural and idiomatic usage in Chinese, not an unnecessary addition of meaning. Comment 1 oversimplifies the issue by treating "也" as a simple semantic addition, ignoring its textual role. Therefore, Comment 2 provides a more accurate and nuanced analysis. DeepSeek agrees with evaluators' disagreement. Evaluators' further investigation confirms the original judgment. DeepSeek is correct.
3	There are two comments on the translation pair below: English: "Asia's integration has come about through commercial pragmatism rather than political alignment." Chinese: "亚洲的融入是通过商业务实性，而不是政治取向一致性。" Comment 1: Linguistic convention issue: "通过商业务实性" is grammatically acceptable but unnatural. "基于/源于" is better. Comment 2: 基于/源于 are better but 通过 isn't wrong and still flies, We wouldn't flag this as an error. Which comment is better? Why?	DeepSeek: Comment 1 is better because it identifies a genuine nuance in idiomatic expression. While "通过" is grammatically acceptable, the English phrase "come about through" implies a causal foundation or driving force—not merely a method or channel. In Chinese, "基于" (based on) or "源于" (deriving from) more naturally convey this underlying reason, making the translation sound more native. Comment 2 is too permissive: it fails to recognize that a small collocation shift can improve clarity and fluency, which is the goal of good translation. Therefore, Comment 1 offers a useful, targeted improvement rather than just excusing a passable but suboptimal choice. DeepSeek does not agree with evaluators' disagreement. Evaluators' further investigation confirm the original judgment. DeepSeek is incorrect.
4	There are two comments on the translation pair below: English: "Can that continue if trade tensions rise?" Chinese: "如果贸易紧张局势加剧，亚洲还可以维持现状吗?" Comment 1: Omission: "if trade tensions rise". Comment 2: There is no omission here. The translation of "Can that continue if trade tensions rise?" is broken down into 2 clauses in Chinese separated by a comma, the translation is "如果贸易紧张局势加	DeepSeek: Comment 2 is better because the Chinese translation provided — "如果贸易紧张局势加剧，亚洲还可以维持现状吗?" — actually includes the conditional clause "if trade tensions rise" (如果贸易紧张局势加剧). Comment 1 incorrectly claims an omission. Comment 2 correctly identifies that the clause is present, placed before the comma, and explains why the earlier alleged omission was mistaken. DeepSeek agrees with evaluators' disagreement. Evaluators' further investigation confirms the original judgment. DeepSeek is correct.

	alleged omission is placed before the comma, the LLM simply missed it. Which comment is better? Why?	Evaluators' further investigation confirms the original judgment. DeepSeek is correct.
5	There are two comments on the translation pair below: English: ""Can that continue if trade tensions rise?" Chinese: "亚洲还可以维持现状吗?" Comment 1: Style (Slightly awkward phrasing) problem to translate "that continue" as "维持现状" Comment 2: The mistranslation seems to be caused by a misunderstanding of the source sentence and it is not the word "continue". Which comment is better? Why?	DeepSeek: Comment 2 is better because it correctly identifies the root cause of the mistranslation: the word "that" in the English source was misunderstood or arbitrarily replaced with "亚洲" (Asia) in the Chinese, which is a semantic error. Comment 1 focuses on a style issue with the translation of "continue" as "维持现状," which is a minor phrasing choice; the real problem is the insertion of "亚洲" that has no basis in the original sentence. Thus, Comment 2 addresses the more serious mistake. DeepSeek agrees with evaluator's disagreement. Evaluators' further investigation confirms the original judgment. DeepSeek is correct.
6	There are two comments on the translation pair below: English: "Asia excels in manufacturing new technology in four areas in particular." Chinese: "亚洲在如下四项新型科技 制造方面超群拔类" Comment 1: Accuracy issue: translate "the following four items" as "如下四项" Comment 2: "以下" sounds better than "如下", but this minor enough that it should not be counted as an error.] Which comment is better? Why?	DeepSeek: Comment 1 is better because it identifies a genuine accuracy issue: the English source does not include the phrase "the following" — it simply says "in four areas in particular." The Chinese translation adds "如下" (as follows), which implies that those areas are about to be listed, a nuance not present in the original. This is an unnecessary and potentially misleading addition. Comment 2, on the other hand, focuses on a minor stylistic preference ("以下" vs. "如下") that does not affect meaning and explicitly states it should not count as an error. Therefore, Comment 1 addresses a more substantive problem with the translation. DeepSeek agrees more with GPT5.4's point. Evaluators' further investigation confirms the original judgment. DeepSeek is incorrect.
7	There are two comments on the translation pair below: English: "Asia has the people to fuel growth, but 90 percent of the expected increase in Asia's non-farm workforces by 2050 will be where productivity is relatively low." Chinese: "亚洲拥有发展所需的人口, 但2050年前亚洲非务农工作力的预期增长的90%将处于生产力相对低的地区。" Comment 1: Terminology issue: translate "non-farm workforces" as "非务农工作力" Comment 2: 非务农 and 非农业 both should work. It is not an error]. Which comment is better? Why?	DeepSeek: Comment 2 is better because it correctly recognizes that both "非务农" and "非农业" are acceptable translations for "non-farm" in this context, and it wisely avoids insisting on a single, less idiomatic phrasing. The term "非农业劳动力" is more standard and widely used in Chinese economic texts, while "非务农劳动力" is also understandable but less common. Additionally, "工作力" in Comment 1 is less natural than the standard "劳动力" for "workforce." Comment 2's open-minded assessment is more accurate and helpful for translation quality. DeepSeek agrees with human disagreements. Evaluators' further

		judgment. DeepSeek is correct.
8	There are two comments on the translation pair below: English: “But now, it could find itself in the crosshairs of trade tensions.”” Chinese: “然而目前亚洲也发现自己紧张的贸易关系中成了靶子。” Comment 1: Accuracy issue: translate “trade tensions” as “紧张的贸易关系”. The source refers specifically to trade tensions/conflicts, which implies geopolitical friction or disputes. The translation weakens the meaning and changes the nuance. Comment 2: The turn of phrase is a bit awkward but it isn’t an error in accuracy. Interpreting it as “conflicts” may be considered an over translation for the exam.] Which comment is better? Why?	DeepSeek: Comment 2 is better. The translation “紧张的贸易关系” (tense trade relations) is a standard and accurate rendering of “trade tensions” — it does not weaken the meaning, as both terms refer to the same geopolitical friction. Comment 1 overstates the issue; there is no significant loss of nuance. Comment 2 correctly identifies the awkwardness in phrasing (the Chinese sentence is grammatically clunky) but recognizes that this is a stylistic flaw, not an accuracy error. Furthermore, Comment 2’s point about over-translation into “conflicts” is valid, as that would be a shift in register. Therefore, Comment 2 offers a more balanced and precise critique. DeepSeek agrees with human disagreements. Evaluators’ further investigation confirms the original judgment. DeepSeek is correct.
9	There are two comments on the translation pair below: English: “But the value created by tech is moving away from hardware and into new transversal technologies. So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” Chinese: “然而创造价值的科技产业正逐渐从硬件转变为新式横向科技，因此问题在于亚洲是否能在科技创新和科技制造业这些领域进行自我重塑。” Comment 1: Accuracy issue to translate “tech creator” as “科技创新” Entire sentence structure is an over-expansion not present in source. Comment 2: The source words “creator” and “manufacturer” are stylistic choice. The “base” meaning of the segment is “create/invent tech” and “manufacture tech”, which is preserved. We would not expect the parts of speech in the translation to align perfectly with the source - syntax works differently in different languages. Which comment is better? Why?	DeepSeek: Comment 1 is better. Reasoning: Comment 1 correctly identifies two specific problems: 1. Accuracy issue: The source uses “tech creator” (an entity that creates technology), but the translation renders it as “科技创新” (technological innovation), which shifts the meaning from a role/agent to an abstract activity. This is a semantic change, not a mere stylistic or syntactic adjustment. 2. Over-expansion: The Chinese adds “这些领域” (these fields) and restructures the sentence into a “reinvent in the fields of X and Y” format, while the source says “reinvent itself as X as well as Y.” This changes the core relationship from identity (being a creator and manufacturer) to domain (operating within innovation and manufacturing fields). Comment 2 attempts to defend the translation by claiming the “base meaning” is preserved and that part-of-speech alignment is unnecessary. However, this overlooks the crucial distinction between being a creator (+ noun, subject) and engaging in innovation (+ abstract noun, field). The translation’s structural expansion also introduces a meaning not present in the source, which goes beyond acceptable syntactic variation. DeepSeek agrees with GPT5.4. Evaluators’ further investigation confirms the original judgment: comment 2 is better. The base meaning analysis in comment 2 is accurate. Comment 1 rigidly separates create(or) and 创新 as two different things, whereas in this context they refer to the

- There are two comments on the translation pair below:
English: "Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies."
Chinese: "亚洲有大批的年轻工人，当中有些经济体的生产力又有所提高，这些在很大程度上都是最有利的人口特征。"
- 10 Comment 1: Accuracy issue: "these largely are" translated as "这些在很大程度上都是" introduces an explanatory structure not present in the source sentence.
Comment 2: We think 很大程度上 matches "largely" in the source text, what's changed is the sentence structure - but the meaning of "largely" is preserved. This is not an error.
Which comment is better? Why?
- DeepSeek:** Comment 2 is better because it correctly recognizes that translation often requires restructuring sentences while preserving meaning. The Chinese version accurately conveys "largely" as "在很大程度上" and the overall idea—that Asia's demographic conditions (young workers and rising productivity) constitute a "demographic sweet spot." The added explanatory structure ("这些...都是") in the Chinese does not distort the source meaning; it simply makes the reference explicit, which is a natural adaptation in Chinese. Comment 1 is overly literal and fails to account for acceptable rephrasing in translation. DeepSeek agrees with human disagreement. **DeepSeek is correct.**
- There are two comments on the translation pair below:
English: "Asia has largely been in a demographic sweet spot with large pools of young workers and surging productivity in some economies."
Chinese: "亚洲从很大程度上一直拥有人口红利，在某些经济体中拥有大量的年轻工人和高涨的生产力。"
- 11 Comment 1: "productivity" is translated as "生产力" instead of the more precise economic term "生产率".
Comment 2: 生产力 is not wrong. Here the paragraph is talking about demographics and labour, and with young workers it naturally factors that the output capability is high. "Productivity" in this context refers more likely to the ability to produce, rather than the ratio of material output over input.
Which comment is better? Why?
- DeepSeek:** Comment 1 is better. While "生产力" (productive forces) is not entirely incorrect, "生产率" (productivity rate/efficiency) is the standard and more precise economic term for "productivity" in this context. The English sentence explicitly separates "large pools of young workers" (a demographic factor) from "surging productivity" (an efficiency factor), so the latter refers to output per worker or per unit of input, not just the general ability to produce. Using "生产率" avoids ambiguity and aligns with common usage in economic translations. DeepSeek agrees with GPT5.4. Evaluators' further investigation confirms the original judgment. **DeepSeek is incorrect.**
- There are two comments on the translation pair below:
English: "But now, aging is fiercest in the high-productivity economies of the Pacific Rim."
Chinese: "但如今，环太平洋的高效经济体中的老龄化问题也最为严重。"
- 12 Comment 1: The word "也" (also) is added in the translation, but the source sentence does not contain the meaning "also."
Comment 2: We can see how 也 may be a problem if one is to back-translate this into English. However, there is an implied contrast between this sentence and the previous sentence in the source. "Asia population situation good. (But) some countries have population problems". 也 is used as an emphasis for the aging issue,
- DeepSeek:** Comment 2 is better because it takes into account the broader context and the pragmatic function of "也" in Chinese. While Comment 1 correctly notes that the word "也" is not directly present in the English source, translation fidelity is not simply about word-for-word equivalence. The source sentence begins with "But now," which signals a contrast with prior discourse. If the preceding text discussed aging in other regions (e.g., "Aging was moderate in other areas"), then using "也" in Chinese can serve to highlight that the same issue is now also present in the Pacific Rim, effectively conveying the contrast and emphasis intended by the source's "But" and "fiercest." In Chinese, "也" often carries such emphatic or contrastive nuance beyond

Which comment is better? Why?

There are two comments on the translation pair below:

English: “but 90 percent of the expected increase in Asia’s non-farm workforces by 2050 will be where productivity is relatively low.”

Chinese: “但在截至2050年的亚洲非农劳动力预期增长中，其中90%的增长都来自生产力相对较低的地区。”

Comment 1: Accuracy issue: “will be where productivity is relatively low” was translated as “来自生产力相对较低的地区”. The source sentence means the workforce increase will occur in low-productivity areas, not that it comes from those areas. “来自” introduces an unintended source meaning.

Comment 2: We think this is fine and would not cause misunderstanding on the part of the reader. The core message refers to where labor growth is, the reader would think about location rather than where the people are born (“from”).

Which comment is better? Why?

There are two comments on the translation pair below:

English: “Asia excels in manufacturing new technology in four areas in particular: consumer electronics, industrial electronics, electric vehicles, and semiconductors”

Chinese: “亚洲在新科技生产方面遥遥领先，尤其在消费电子产品、工业电子、电动汽车和半导体四个方面。”

Comment 1: Accuracy issue: “consumer electronics” is translated as “消费电子产品”. Adding “产品” slightly expands the meaning. While common in Chinese, it is not strictly present in the source text.

Comment 2: Such term exists and is used in Chinese. It does not lead to any misunderstanding.

Which comment is better? Why?

There are two comments on the translation pair below:

English: “So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.”

Chinese: “如此一来，问题就变成亚洲是否能不仅作为技术生产者，也是技术创新者进行再次创新。”

Comment 1: Addition: The phrase “如此一来” (“thus/therefore”) is added but does

overlooks these pragmatic considerations and imposes a overly literal standard, whereas Comment 2 demonstrates a more nuanced understanding of translation as conveying meaning and discourse cohesion.

DeepSeek agrees with human evaluator. Evaluators’ further investigation confirms the original judgment. **DeepSeek is correct.**

DeepSeek: Comment 2 is better because it recognizes that the translation is idiomatic and does not cause misunderstanding. The Chinese phrase “来自...地区” is commonly used to indicate where something occurs or is located (e.g., “90% of the increase comes from low-productivity areas” naturally means the increase happens in those areas), not necessarily implying origin or birthplace. The core message—that the workforce growth will happen in low-productivity regions—is preserved. Comment 1 over-interprets the preposition “来自” and applies a stricter, less context-sensitive standard that would make many natural translations seem inaccurate.

DeepSeek agrees with human evaluator. Evaluators’ further investigation confirms the original judgment. **DeepSeek is correct.**

DeepSeek: Comment 2 is better. The addition of “产品” (products) in the Chinese translation is natural and idiomatic in Chinese. “Consumer electronics” (消费电子产品) is a standard, widely accepted term in Chinese that accurately conveys the intended meaning. The slight expansion does not create any misunderstanding or distort the original message, and it aligns with common usage rather than being a strict accuracy error. Comment 1 is overly pedantic, as it fails to account for natural target-language conventions.

DeepSeek agrees with human evaluator. Evaluators’ further investigation confirms the original judgment. **DeepSeek is correct.**

DeepSeek: Comment 2 is better because it goes beyond mere observation to evaluate the translation's appropriateness. It correctly recognizes that “如此一来” is a natural Chinese discourse marker that fulfills the same cohesive function as “So” in the source—linking the statement to previous context. Comment 1 simply notes an addition without considering

is wrong. It is likely serves the function “so” serves in the source. “Given the above context” is the idea “so” and 如此一来 express. The translation is appropriate.]

translation, missing the key principle that good translation often requires adding or omitting elements for naturalness and clarity in the target language.

Which comment is better? Why?

DeepSeek agrees with human evaluator. Evaluators’ further investigation confirms the original judgment. **DeepSeek is correct.**

Table 7: DeepSeek arbitrated disagreements on negative exemplars between GPT5.4 and human evaluators. Based on the human evaluators’ further investigation and final judgments, DeepSeek’s arbitration was correct in 11 of the 15 cases.

#	GPT5.4 generated explanation (comment 1) and human evaluator’s disagreements (comment 2)	DeepSeek’s arbitration and Human evaluator’s Re-evaluation
1	There are two comments on the translation pair below: English: “We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.” Chinese: “我们相信世界正处于新时代的开端，与过去30年相比，影响世界的根本势力已变得非常不同。” Comment 1: “underlying force” is translated as “根本势力”: The phrase “very different underlying forces” is effectively translated as “根本势力已变得非常不同,” preserving the conceptual emphasis on foundational drivers of change. There is no omission or distortion of meaning. Comment 2: Translating “underlying” as 根本 is not wrong enough to warrant a point deduction, but 根本 can mean “source/origin” whereas the meaning of “underlying” is more like “deep down/deep within”. Therefore, this is a minor false positive.] Which comment is better? Why?	DeepSeek: Comment 2 is better because it offers a more nuanced and accurate linguistic critique. While Comment 1 states that the translation preserves meaning without distortion, Comment 2 correctly identifies a subtle semantic shift: “根本” leans toward “source/origin,” whereas “underlying” conveys “deep down / deep within.” This distinction matters in professional translation evaluation, as it reflects a slight misalignment in connotation. Comment 2 is also more helpful because it acknowledges the translation is acceptable (no point deduction) while still highlighting a minor false positive, thus providing balanced and precise feedback. Comment 1 oversimplifies the issue by claiming no distortion at all. DeepSeek agrees with human evaluator. Evaluators’ further investigation confirms the original judgment. DeepSeek is correct.
2	There are two comments on the translation pair below: English: “So, the question is whether Asia can reinvent itself as a tech creator as well as a tech manufacturer.” Chinese: “因此问题在于亚洲是否能在科技创新和科技制造业这些领域进行自我重塑。” Comment 1: “reinvent itself”→“自我重塑”: the notion of “reinvent itself” is effectively rendered as “进行自我重塑,” which captures the idea of transformation rather than mere adjustment. Comment 2: 重塑 is used well, but 自我 personifies the subject Asia too much (not intended in source text). Which comment is better? Why?	DeepSeek: Comment 1 is better because it accurately recognizes that the source text intentionally personifies Asia through the reflexive pronoun “itself” in “reinvent itself.” The Chinese translation “进行自我重塑” faithfully preserves this personification, as the metaphor of reinvention inherently implies agency. Comment 2 incorrectly claims the personification is unintended, when in fact the English explicitly uses “itself” to personify Asia, making “自我” an appropriate and natural rendering. DeepSeek agrees with GPT5.4. Evaluators’ further investigation (自我重塑 isn’t wrong but it personifies Asia (as one entity). Asia does not act as one entity, what the writer means is whether the countries in Asia can make a shift) confirms the original judgment. DeepSeek is incorrect.

pair below:

English: "We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years."

Chinese: "我们认为全球正处在一个新时代的风口浪尖, 各种深层力量相互角逐。而这几股力量与过去30年间塑造世界的力量截然不同。"

Comment 1: "the cusp of a new era" → "新时代的风口浪尖": The core meaning of the source sentence is conveyed clearly and faithfully. The notion of "on the cusp of a new era" is effectively captured by "正处在一个新时代的风口浪尖," which reflects a moment of transition and significance.

Comment 2: The idiom is good (so the comment about it being lively isn't wrong). But it does not accurately reflect "cusp". "Cusp" means "at the threshold" or "change is about to happen", the segment about the underlying forces describes the subject of "on the cusp". 风口浪尖 is more on the side or "at the height of tensions"; it signifies underlying struggle, which is not suggested in the source text and therefore distorts the meaning.

Which comment is better? Why?

There are two comments on the translation pair below:

English: "Asia has the people to fuel growth, but 90 percent of the expected increase in Asia's non-farm workforces by 2050 will be where productivity is relatively low."

Chinese: "亚洲虽然拥有人力来推动增长, 但在截至2050年的亚洲非农劳动力预期增长中, 其中90%的增长都来自生产力相对较低的地区。"

4 Comment 1: "90 percent of the expected increase" is translated as "其中90%的增长":

The translation reads naturally and fluently in Chinese. The use of "截至2050年" and "其中90%的增长都来自..." creates a smooth, idiomatic flow rather than a rigid, word-for-word rendering. The sentence maintains a formal, analytical tone consistent with the original text.

Comment 2: Neutral, the translation is correct, but nothing outstanding, and using 增长 twice in close succession may seem redundant to some.

Which comment is better? Why?

Deep Seek: Comment 2 is better because it identifies a significant semantic inaccuracy in the translation. The English phrase "on the cusp of a new era" means being at the threshold or on the verge of a new era, emphasizing imminent change—not being in the midst of a struggle. The Chinese idiom "风口浪尖" (feng kou lang jian) literally means "at the mouth of the wind and the tip of the waves," and figuratively refers to being at the center of a storm or in a highly tense, contentious situation. This connotation of active conflict or turmoil is absent from the original, which simply states that the underlying forces are different, not that they are locked in competition. Additionally, the translation gratuitously adds "各种深层力量相互角逐" ("various deep forces are competing with each other"), which further distorts the original meaning. Comment 1 mistakenly equates "cusp" with "风口浪尖" by focusing only on the transitional sense, ignoring that the idiom carries a strong implication of struggle that misrepresents the source text. Thus, Comment 2 provides a more accurate, linguistically informed critique.

DeepSeek agrees with human evaluators. Evaluators' further investigation confirms the original judgment. **DeepSeek is correct.**

DeepSeek: Comment 1 is better. It provides a detailed, positive assessment of the translation's natural fluency and idiomatic structure, highlighting specific elements like the handling of the temporal phrase "by 2050" and the use of "其中...都来自" to create a smooth flow. This analysis demonstrates a deeper understanding of translation quality, focusing on overall effectiveness rather than a minor stylistic quibble.

Comment 2, while valid in noting the repeated use of "增长," is overly neutral and focuses on a potential redundancy that does not significantly detract from the translation's correctness or readability. In Chinese, such repetition is often acceptable for clarity and emphasis, making this critique less impactful. A good translation critique should weigh strengths against minor weaknesses, and Comment 1 does that more insightfully.

DeepSeek agrees with GPT5.4. Evaluators' further investigation confirms the original judgment.

DeepSeek is correct.

DeepSeek: Comment 2 is better because it prioritizes semantic fidelity over stylistic embellishment. The

English: “We believe that the world is on the cusp of a new era with very different underlying forces than those that shaped our world over the past 30 years.”

Chinese: “人们认为相较于过去30年塑形世界的力量，新时代截然不同的潜在力量，将世界推向风口浪尖。”

Comment 1: “the cusp of a new era “□“新时代的风口浪尖” : The sentence reads smoothly and idiomatically in Chinese. Expressions like “风口浪尖” effectively convey a sense of urgency and transformation, making the sentence engaging and vivid rather than stiff or overly literal.

Comment 2: Translation can be subjective. There are readers who marvel at the use of fancy words without thinking deeper at the meaning. But We think here since the meaning is not fully reflective, the negative outweighs the positive.] Which comment is better? Why?

forward-looking perspective (“on the cusp of a new era,” “We believe”), while the Chinese translation introduces the dramatic, crisis-laden idiom “风口浪尖” (at the crest of a wave / in the teeth of a storm) and changes the subject from “We” to “人们” (people). This alters both the tone and the agency. Comment 1 merely praises the vividness without addressing these inaccuracies, which is a superficial evaluation for a translation task. Comment 2 correctly identifies that the meaning is not fully reflected, and that the negative impact of distorting the original outweighs any stylistic gains. A good translation must first be faithful; flair is secondary.

DeepSeek with human evaluators. Evaluators’ further investigation confirms the original judgment. **DeepSeek is correct.**

There are two comments on the translation pair below:

English: “But the value created by tech is moving away from hardware and into new transversal technologies.”

Chinese: “然而技术产生的价值正从硬件生产转移到新兴交叉科技。”

Comment 1: “transversal technologies” is translated as “交叉科技”: The rendering of “transversal technologies” as “交叉科技” is particularly strong. It reflects an understanding of the concept as cross-domain or interdisciplinary technologies, which is consistent with how such ideas are expressed in Chinese technical and business contexts.

Comment 2: this term does not seem adopted in Chinese and does not make much sense at face value.

Which comment is better? Why?

DeepSeek: Comment 1 is better because it accurately recognizes that “交叉科技” is a valid and idiomatic translation of “transversal technologies” in Chinese technical and business discourse. The term conveys the idea of cross-domain or interdisciplinary technologies, which aligns with common usage in Chinese innovation contexts. Comment 2’s assertion that the term is not adopted and makes little sense is incorrect—it is actually established in relevant literature.

DeepSeek agrees with GPT5.4. Evaluators’ further investigation confirms the original judgment (交叉科技 does not make sense to native Chinese speaker). **DeepSeek is incorrect.**

There are two comments on the translation pair below:

English: “But now, aging is fiercest in the high-productivity economies of the Pacific Rim.”

Chinese: “可是如今，老龄化在太平洋地区高生产力的经济体中成为最棘手的问题。”

Comment 1: The core meaning of the source sentence is well preserved. “aging is fiercest” is effectively rendered as “老龄化...成为最棘手的问题,” which captures the intensity and seriousness of the demographic challenge, even though it is slightly interpretive. The relationship between aging and “high-productivity economies of the Pacific Rim” is clearly and correctly conveyed.

Comment 2: The translation is idiomatic, but it also distorts the meaning.

DeepSeek: Comment 1 is better because it accurately assesses the translation’s strengths and acknowledges its slight interpretive nature without overstating the flaws. Why Comment 1 is more accurate: The translation “老龄化在太平洋地区高生产力的经济体中成为最棘手的问题” indeed captures the core meaning of the source: it conveys that aging is a severe and pressing issue in the specified economies. The shift from “is fiercest” (suggesting intensity) to “成为最棘手的问题” (becoming the most troublesome problem) is a natural idiomatic adjustment in Chinese. It preserves the essence—that aging is at its most acute in these regions—while sounding natural in the target language. Comment 1 correctly notes this adaptation as “slightly interpretive” but

Comment 2 claims the translation “distorts the meaning,” which is too strong. While there is a minor nuance change (from a description of intensity to a problem-oriented phrasing), the overall meaning—that aging is most severe in high-productivity Pacific Rim economies—is not distorted. The translation does not introduce false information or contradict the source; it simply uses a different but equivalent expression for “fiercest” in Chinese. Therefore, calling it a distortion is an overstatement and fails to recognize the acceptable degree of idiomatic adaptation.

In summary: Comment 1 provides a balanced, context-aware evaluation; Comment 2 is overly critical and not fully justified by the actual translation. DeepSeek agrees with GPT5.4. Evaluators’ further investigation confirms the original judgment
DeepSeek is incorrect.

Table 8: DeepSeek arbitrated disagreements on positive exemplars between GPT5.4 and human evaluators. Based on the human evaluators' further investigation and final judgments, DeepSeek's arbitration was correct in 3 cases out of 7.

Predict and Fix: A Unified Model for Translation Quality Estimation and Post-Editing

Maciej Modrzejewski

BIG Language Solutions

maciej.modrzejewski@biglanguage.com

Yash Bhaskar

BIG Language Solutions

yash.bhaskar@biglanguage.com

Chinmay Pateria

BIG Language Solutions

chinmay.pateria@biglanguage.com

Abstract

We propose a unified architecture for jointly modeling Translation Quality Estimation (QE) and Automatic Post-Editing (APE) within a single lightweight language model. Our approach integrates quality prediction and correction generation in a single decoding process using a decoder-only Qwen2.5 model (0.5B parameters), augmented with a dedicated QE regression head operating on hidden states at a special token position. The model produces structured outputs that include a continuous quality score, an edit decision, and a corrected translation when necessary. We train on datasets of 100K, 1M, and 1.84M manually annotated samples across eight language pairs, enabling analysis of both data scale and distribution. Experimental results show that the proposed model achieves strong QE performance ($r = 0.907$) and high post-editing decision accuracy (88.4%), while reducing over-editing compared to both autoregressive baselines and large commercial LLMs.

1 Introduction

Quality Estimation (QE) provides reference-free assessment of machine translation (MT) output and is widely used in localization workflows to determine whether a translation can be published as-is or should be revised (Zerva et al., 2024). When revision is required, the segment is forwarded to either a human post-editor or an Automatic Post-Editing (APE) system, which in recent deployments is typically a large language model (LLM), often combined with Retrieval-Augmented Generation (Lewis et al., 2020) for terminology and domain conditioning.

In this conventional setup, QE and APE are implemented as two independent models invoked sequentially. The QE model returns a scalar quality score, a thresholding rule converts the score into a routing decision, and a separate APE model produces the corrected output. This decoupling has two consequences. First, the internal representations that the QE model uses to detect translation errors are dis-

carded before post-editing begins, so the APE model must recover comparable information from the same input (Ranasinghe et al., 2020). Second, two forward passes through large models are required for every segment.

We study whether QE and APE can be performed jointly within a single decoder-only model, such that a single forward pass yields (i) a continuous quality score, (ii) PE action and (iii) a corrected translation when PE is needed. The model is built on Qwen2.5-0.5B (Qwen et al., 2025) and is extended with a regression head attached to the hidden state at a designated $\langle \text{QE} \rangle$ token position whereas the language modeling head produces the edit decision and post-edited text. The two objectives are optimized together using an uncertainty-weighted multi-task loss (Kendall et al., 2018).

We evaluate this design on manually annotated QE+APE data covering eight English-source language pairs at three training scales (100K, 1M, and

1.84M segments). The contributions of this work are:

- A unified decoder-only architecture that produces a calibrated QE score, an edit decision, and a post-edited translation in a single decoding pass.
- An empirical comparison against commercial LLMs and a decoupled QE+APE pipeline across eight language pairs, including an analysis of over and forced editing behavior.
- A study on the effect of scaling training data and distribution of language pairs on joint QE+APE performance.

2 Related Work

2.1 Reference-free Quality Estimation

Sentence-level QE has been dominated by learned metrics built on multilingual pre-trained encoders. COMET (Rei et al., 2020a) introduced a regression framework over cross-lingual representations, and TransQuest (Ranasinghe et al., 2020) showed that fine-tuned cross-lingual transformers showed gains in representations effectively for QE across language pairs. Subsequent systems such as xCOMET (Guerreiro et al., 2023) and MetricX-23 (Juraska and Others, 2023) extend this line with finer-grained error supervision and report strong correlation with human judgments at the WMT shared tasks (Zerva et al., 2024). In parallel, prompt-based methods including GEMBA (Kocmi et al., 2023; Kocmi and Federmann, 2023) obtain quality assessments directly from large language models. These approaches produce a quality score but do not generate a corrected translation, so any downstream revision still requires a separate model.

2.2 LLM-based Automatic Post-Editing

A second line of work uses large language models for direct post-editing of MT outputs (Zhang et al., 2023; Raunak and Others, 2023; Brown et al., 2020). Such systems are typically prompted in a zero or few shot manner with the source segment and the MT hypothesis and asked to return a revised translation. Variants include error-annotation guided post-editing (Ki and Carpuat, 2024), iterative refinement (Chen and Others, 2023), and self-refinement (Feng and Others, 2024). A repeated observation in this literature is that generative post-editors lack a calibrated decision

signal for whether an edit is necessary, which leads to unnecessary modifications of acceptable translations, Huang et al. (2024) further report that LLMs do not reliably self-correct without an external signal.

2.3 Joint and Quality-informed Post-Editing

WMT 2024 Quality-informed Automatic Post-Editing track shared tasks have explicitly targeted the integration of QE and APE also it required systems to combine quality prediction with correction, and quality-aware decoding methods. Most existing systems still couple the two stage process through prompting or pipeline orchestration rather than a unified approach. In contrast, this work performs QE and APE jointly inside a single decoder-only model, with quality estimation produced by a regression head over a designated token. In the same decoding pass the post-edited output is produced, and evaluates the design across eight language pairs and three training scales.

3 Methodology

3.1 Problem Formulation

We formulate joint Quality Estimation and Post-Editing as a single conditional prediction problem. Given a source sentence s and an MT hypothesis m_0 , the model jointly predicts a quality score $\hat{q} \in [0, 1]$, a binary edit decision $a \in \{\text{No PE}, \text{PE}\}$, and a post-edited translation f (defined only when $a = \text{PE}$):

$$(\hat{q}, a, f) = \mathcal{F}_\theta(s, m_0). \quad (1)$$

All three outputs are obtained from a single forward pass of the same model, without additional prompting steps or external thresholding (Figure 1).

The QE supervision signal is derived from the normalized edit distance (NED) between the MT output and the linguist correction:

$$\text{NED} = \frac{\text{Lev}(\text{MT}, \text{PE})}{\max(|\text{MT}|, |\text{PE}|)}, \quad q = 1 - \text{NED}, \quad (2)$$

clipped to $[0, 1]$. Higher q corresponds to fewer required edits.

We use NED as a practical proxy for sentence-level QE because it directly measures the amount of human post-editing required to transform the MT hypothesis into the approved linguist correction. In localization workflows, this edit effort is closely aligned with translation quality: outputs requiring

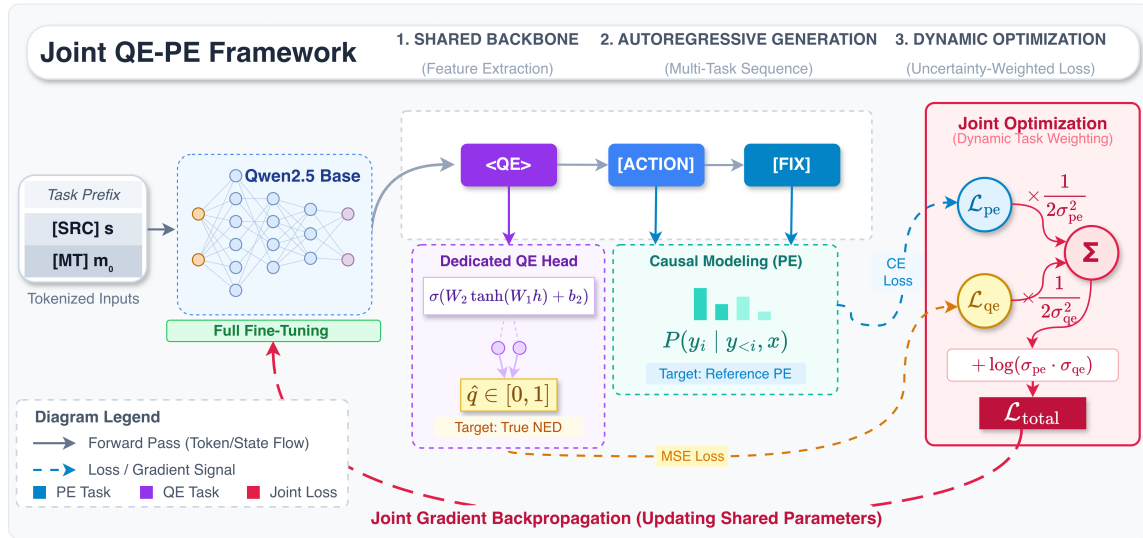


Figure 1: Joint QE+APE architecture. Given a source sentence and an MT hypothesis, the Qwen2.5 backbone autoregressively decodes a single output sequence containing the <QE> token, an edit decision (No PE / PE), and a post-edited translation when applicable. The final-layer hidden state at the <QE> position is read once and passed through a regression head to produce the scalar quality score $\hat{q}$. The language modeling and regression objectives are optimized jointly with uncertainty-weighted losses; no additional model calls or external thresholding are required at inference.

fewer edits are generally more acceptable for publication, whereas outputs requiring substantial edits indicate lower quality. Although NED does not capture all semantic adequacy or fluency issues perfectly, it provides a consistent, reference-based supervision signal that is available for every post-edited segment and is therefore suitable for training a calibrated QE regression model.

3.2 Architecture

The backbone is a decoder-only language model in our experiments we use Qwen2.5-0.5B (Qwen et al., 2025), but the formulation applies to any autoregressive transformer. The vocabulary is extended with a special token <QE>, following the use of dedicated control tokens in prior quality-aware work (Koneru et al., 2025). During training, the input sequence is constructed so that the <QE> token precedes the structured output containing the edit decision and the post-edited translation. The hidden state at the position of this token is used as the input to the QE regression head.

Let $h_{\langle QE \rangle} \in \mathbb{R}^d$ denote the final-layer hidden state at the special token <QE> position. The QE

score is obtained by a two-layer regression head:

$$\hat{q} = \sigma(\mathbf{W}_2 \cdot \text{Dropout}(\tanh(\mathbf{W}_1 h_{\langle QE \rangle} + b_1)) + b_2), \quad (3)$$

with $\mathbf{W}_1 \in \mathbb{R}^{d \times d}$, $\mathbf{W}_2 \in \mathbb{R}^{1 \times d}$. The sigmoid output constrains $\hat{q}$ to $[0, 1]$ to match the NED-derived target. The language modeling head of the main pre-trained model is unchanged and produces the structured output token sequence containing the action label and, when applicable, the corrected translation.

Because the regression head consists of only two linear layers operating on an already-computed hidden state, it adds negligible computational overhead during inference.

3.3 Training Objective

The model is trained with two losses computed on the same forward pass: a regression loss $\mathcal{L}_{qe}$ (mean squared error between $\hat{q}$ and the NED-derived target) and the standard cross-entropy language modeling loss $\mathcal{L}_{\text{text}}$ over the structured output tokens. The two losses are combined using homoscedastic uncertainty weighting (Kendall et al., 2018), with learnable log-variance parameters s_{qe} and s_{text} :

$$\mathcal{L}_{\text{joint}} = e^{-s_{qe}} \mathcal{L}_{qe} + e^{-s_{\text{text}}} \mathcal{L}_{\text{text}} + s_{qe} + s_{\text{text}}. \quad (4)$$

The exponential representation keeps the task weights positive, and the additive s_{qe} and s_{text} terms prevent the loss from being minimized by increasing the log-variances without bound. Both s_{qe} and s_{text} are updated jointly with the model parameters during training.

4 Experimental Setup

4.1 Data

We construct three training datasets of approximately 100K, 1M, and 1.84M segments covering eight English-source language pairs: French (Canadian), Japanese, Spanish, Korean, Simplified Chinese, German, Polish, and European French. For every segment, the MT hypothesis is produced by an in-house translation system fine-tuned for the target domain, and the reference is a post-edited correction approved by professional linguists (gold references). The 100K and 1M datasets follow the natural distribution of the source corpus, whereas the 1.84M dataset is rebalanced to allocate approximately 250K segments per language pair, with the exception of English to European French (93,364 segments), for which fewer instances were available. The per-pair distribution for each scale is reported in Table 1.

For evaluation, we hold out a test set of 9,048 segments sampled from the same distribution as the training data. Table 2 reports BLEU and COMET scores between the in-house MT output and the gold references on this test set. Scores are high across pairs (BLEU above 60 on most pairs, COMET close to or above 1.0 on French, Spanish, and Korean), which limits the room for surface-metric improvement through post-editing on the higher-resource pairs.

Each segment is annotated with a binary decision (PE or No PE) and a corrected translation when applicable. The continuous regression target is $q = 1 - \text{NED}$, where NED is the length-normalized Levenshtein distance between the MT hypothesis and the gold reference (Section 3).

4.2 Baselines

We compare the proposed joint model against three baselines. The first is an autoregressive variant of our own model, and the remaining two are zero-shot commercial LLMs. All three are evaluated on the same 9,048 segment test set.

Autoregressive QE+APE. The same Qwen2.5-

0.5B model is fine-tuned without the regression head. The QE score is generated as part of the output sequence rather than predicted from the $\langle \text{QE} \rangle$ hidden state. This baseline isolates the contribution of the regression head when training data is the same. We train this baseline on the 100K split.

GPT-5.4 and Gemini 3.1 Pro are evaluated in a zero-shot setting. Since we cannot access the hidden state of these models for the regression head, we prompt these models to output a numeric score with an action and a corrected translation if applicable. The full prompt used here is shown below.

Zero-Shot LLM Baseline Prompt

SYSTEM: You are a translation quality judge from {src_lang} into {tgt_lang}.

Given only a source sentence (SOURCE) and a machine translation (MT), decide if any post-editing (PE) is needed.

Assign a QE score between 0 and 100 where:

100 = perfect translation, no errors

0 = completely incorrect translation.

If no PE is needed, output exactly:

QE: <score between 0 and 100>

ACTION: No Post-Edition

If Post-Edition is needed, output exactly:

QE: <score between 0 and 100>

ACTION: Post-Edition

FIX: <the corrected translation>

Always return only these fields in this order.

USER:

[LANGPAIR] {src_lang}-{tgt_lang}

[SOURCE] {src_text}

[MT] {mt_text}

Language Pair	BLEU	COMET	#Samples
eng→fr_ca	85.62	1.062	3706
eng→jap	69.21	0.808	1067
eng→spa	82.56	1.046	1001
eng→kor	77.18	0.991	942
eng→chi_s	66.13	0.757	906
eng→ger	69.17	0.793	732
eng→pol	62.85	0.978	576
eng→fre	61.15	0.869	118

Table 2: Baseline MT quality on the test set across language pairs, measured in BLEU and COMET, with the number of test samples per pair.

Language Pair	100K %	100K Segments	1M %	1M Segments	1.84M %	1.84M Segments
eng→fr_ca	36.78%	36,780	36.69%	366,900	13.56%	250,000
eng→jap	10.99%	10,990	11.03%	110,300	13.56%	250,000
eng→spa	10.33%	10,330	10.31%	103,100	13.56%	250,000
eng→kor	9.82%	9,820	9.86%	98,600	13.56%	250,000
eng→chi_s	8.84%	8,840	8.80%	88,000	13.56%	250,000
eng→ger	7.73%	7,730	7.69%	76,900	13.56%	250,000
eng→pol	6.13%	6,130	6.08%	60,800	13.56%	250,000
eng→fre	1.22%	1,220	1.24%	12,400	5.06%	93,364
Total		100,000		1,000,000		1,843,364

Table 1: Distribution of language pairs across the 100K, 1M, and 1.84M datasets with corresponding segment counts.

Metric	GPT-5.4	Gemini 3.1 Pro
Global QE Performance		
Pearson r	0.2363	0.3096
Spearman ρ	0.3706	0.3314
MAE	0.1656	0.1515
RMSE	0.2775	0.2599
Global PE Decision		
Accuracy	0.6584	0.6132
Precision	0.7427	0.7972
Recall	0.6805	0.4976
F1 Score	0.7102	0.6127

Table 3: Global QE and PE decision performance for commercial LLM baselines.

Unified QE+APE Prompt (Ours)

SYSTEM: You are a translation quality judge from {src_lang} into {tgt_lang}.

Given only a source sentence (SOURCE) and a machine translation (MT), decide whether post-editing (PE) is needed.

If no PE is needed, output exactly:

QE: <QE>

ACTION: No Post-Edition

If PE is needed, output exactly:

QE: <QE>

ACTION: Post-Edition

FIX: <the corrected translation>

Always return only these fields in this order.

USER:

[LANGPAIR] {src_lang}-{tgt_lang}

[SOURCE] {src_text}

[MT] {mt_text}

4.3 Proposed Architecture

All fine-tuning experiments for the proposed method uses Qwen2.5-0.5B (Qwen et al., 2025) model, with a regression head described in Section 3 attached to the hidden state at the <QE> token position. The model is trained with the structured output prompt shown below, which fixes the <QE> token in the output.

4.4 Evaluation Metrics

We report three groups of metrics.

Quality Estimation. We compute Pearson’s r and Spearman’s ρ between predicted QE scores and the gold scores. We also report Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) between them.

PE Decision Classification. For the PE or No PE prediction, we report Accuracy, Precision, Recall, and F1. We additionally report the accuracy on the No PE subset, which measures how often the model reports no PE required for an acceptable translation.

PE Quality. For segments on which the model decides to PE, we measure the Δ BLEU and Δ COMET (Unbabel/wmt20-comet-da) (Rei et al., 2020b) for translation quality relative to the original MT output.

5 Results and Analysis

We organize the analysis around four research questions: (i) Does the QE regression head improve quality prediction over an autoregressive QE baseline trained on the same data? (ii) How do QE and APE performance scale with training set size and language-pair balancing? (iii) How does the proposed model compare with zero-shot commercial LLMs? (iv) How does post-editing quality vary

Lang Pair	GPT-5.4				Gemini 3.1 Pro			
	PE Acc	QE r	Δ BLEU	Δ COMET	PE Acc	QE r	Δ BLEU	Δ COMET
eng→fr_ca	0.656	0.293	-15.25	-0.016	0.616	0.281	-20.27	-0.059
eng→jap	0.658	0.223	-1.46	-0.006	0.577	0.266	+6.76	-0.024
eng→spa	0.618	0.124	-17.69	-0.021	0.613	0.338	-23.10	-0.059
eng→kor	0.679	0.337	-11.85	-0.001	0.623	0.423	-7.42	-0.021
eng→chi_s	0.648	0.251	+6.84	+0.004	0.622	0.315	+5.56	-0.011
eng→ger	0.677	0.310	-10.74	-0.013	0.659	0.336	-15.90	-0.047
eng→pol	0.693	0.253	-11.05	-0.003	0.610	0.302	-10.80	-0.042
eng→fre	0.695	0.347	-4.10	-0.019	0.441	0.099	-21.50	-0.190

Table 4: Per-language zero-shot post-editing performance for GPT-5.4 and Gemini 3.1 Pro. Δ BLEU and Δ COMET indicate the change relative to the original MT output.

Model	Pearson	Spearman	MAE	RMSE	Acc (All)	Acc (No PE)	Acc (PE)
Baseline AR (100K)	0.094	0.076	0.357	0.528	0.798	0.689	0.866
QE+APE (100K)	0.809	0.762	0.078	0.152	0.822	0.751	0.866
QE+APE (1M)	0.863	0.837	0.059	0.125	0.877	0.814	0.914
QE+APE (1.84M)	0.907	0.876	0.054	0.112	0.884	0.777	0.939

Table 5: QE regression performance and APE decision accuracy. Accuracy (All) is the overall binary decision of whether post-editing is needed. Acc (No PE) measures correct identification of sentences that do not require post-editing, while Acc (PE) measures correct identification of sentences that do require post-editing.

across language pairs?

5.1 Effect of the QE Regression Head

Table 5 compares the autoregressive baseline trained on 100K segments with the proposed QE+APE model trained on the same amount of data. Replacing in-sequence QE score generation with a regression head over the $\langle$ QE $\rangle$ hidden state substantially improves QE performance. Pearson’s r increases from 0.094 to 0.809, while RMSE decreases from 0.528 to 0.152.

The regression head also improves PE decision quality. At the same 100K scale, overall PE decision accuracy increases from 79.8% for the autoregressive baseline to 82.2% for the proposed model. The subset accuracies show that the autoregressive baseline achieves 68.9% accuracy on No-PE examples and 86.6% on PE examples, indicating a tendency to over-edit acceptable translations. The proposed model improves No-PE recognition to 75.1% while maintaining the same PE accuracy of 86.6%, suggesting that explicit QE supervision helps the model make more selective post-editing decisions.

5.2 Effect of Training Scale and Language Balance

Increasing the training data improves both QE prediction and PE decision accuracy. As shown in Table 5, Pearson’s r increases from 0.809 with 100K samples to 0.863 with 1M samples and 0.907 with 1.84M samples. Similarly, overall PE decision accuracy increases from 82.2% to 87.7% and then to 88.4%. MAE and RMSE also decrease consistently, showing that larger training sets produce better calibrated QE scores.

The per-language post-editing results in Table 6 show a similar trend. On Japanese, Korean, German, Polish, and Simplified Chinese, the larger models generally produce stronger Δ BLEU and Δ COMET results. In several cases, scaling turns negative or weak post-editing outcomes into positive gains. For example, German improves from -15.37 Δ BLEU with the autoregressive 100K baseline to $+6.17$ with the 1.84M QE+APE model, while Polish improves from -12.57 to $+0.69$.

However, the effect of scale is also influenced

Lang Pair	Baseline AR (100K)		QE+APE (100K)		QE+APE (1M)		QE+APE (1.84M)	
	ΔBLEU	ΔCOMET	ΔBLEU	ΔCOMET	ΔBLEU	ΔCOMET	ΔBLEU	ΔCOMET
eng→fr_ca	+5.10	+0.03	-0.55	-0.01	+8.09	+0.08	+1.49	-0.02
eng→jap	+8.86	+0.24	+5.03	+0.14	+7.65	+0.20	+7.32	+0.17
eng→spa	-20.04	-0.18	-19.57	-0.14	-16.50	-0.15	-15.64	-0.18
eng→kor	+1.23	-0.03	+5.49	+0.06	+15.19	+0.08	+16.72	+0.09
eng→chi_s	+3.71	+0.07	+4.90	+0.12	+5.13	+0.13	+6.68	+0.07
eng→ger	-15.37	-0.15	-9.85	-0.17	-2.30	-0.08	+6.17	+0.02
eng→pol	-12.57	-0.34	-15.27	-0.23	-2.67	-0.09	+0.69	+0.02
eng→fre	-19.62	-0.47	-18.47	-0.32	-10.39	-0.19	-12.11	-0.18

Table 6: Average change in BLEU and COMET after post-editing relative to the original MT output for each language pair and model configuration. Positive values indicate improvements over the original MT, while negative values indicate degradation. **Green** highlights the best-performing configuration for a given language pair and metric, whereas **red** marks the worst-performing (most negative) result.

Lang Pair	GPT-5.4		Gemini 3.1 Pro		QE+APE (1M) [Ours]		QE+APE (1.84M) [Ours]	
	ΔBLEU	ΔCOMET	ΔBLEU	ΔCOMET	ΔBLEU	ΔCOMET	ΔBLEU	ΔCOMET
eng→fr_ca	-15.25	-0.016	-20.27	-0.059	+8.09	+0.080	+1.49	-0.020
eng→jap	-1.46	-0.006	+6.76	-0.024	+7.65	+0.200	+7.32	+0.170
eng→spa	-17.69	-0.021	-23.10	-0.059	-16.50	-0.150	-15.64	-0.180
eng→kor	-11.85	-0.001	-7.42	-0.021	+15.19	+0.080	+16.72	+0.090
eng→chi_s	+6.84	+0.004	+5.56	-0.011	+5.13	+0.130	+6.68	+0.070
eng→ger	-10.74	-0.013	-15.90	-0.047	-2.30	-0.080	+6.17	+0.020
eng→pol	-11.05	-0.003	-10.80	-0.042	-2.67	-0.090	+0.69	+0.020
eng→fre	-4.10	-0.019	-21.50	-0.190	-10.39	-0.190	-12.11	-0.180

Table 7: Language-level comparison of post-editing impact (ΔBLEU and ΔCOMET). Best results per metric are in **bold**.

by language-pair distribution. The 1.84M dataset is more balanced across language pairs, which improves performance for several lower-resource pairs. At the same time, Canadian French receives a smaller relative share in the 1.84M setting than in the naturally distributed 1M setting, and the 1M model performs better for this pair. This suggests that scaling helps most when it increases both total data volume and sufficient per-language coverage.

5.3 QE and APE Prediction Quality/Accuracy

Table 8 compares our Qwen-based QE+APE models trained on 1M and 1.84M samples against zero-shot GPT-5.4 and Gemini 3.1 Pro. The commercial LLMs obtain PE decision accuracies of 65.8% and 61.3%, respectively, with relatively low QE correlations ($r = 0.236$ and $r = 0.310$). In contrast, the

Qwen QE+APE models are substantially stronger: the 1M model reaches 87.7% PE decision accuracy and $r = 0.863$, while the 1.84M model further improves to 88.4% accuracy and $r = 0.907$.

These results show that the proposed Qwen models provide much more reliable QE and APE predictions than the baseline LLMs. The improvement suggests that task-specific joint training with an explicit QE regression head is more effective for calibrated translation-quality prediction than zero-shot prompting of larger general-purpose LLMs.

5.4 Per-Language Post-Editing Quality

Tables 6 and 7 report ΔBLEU and ΔCOMET on the segments where each model selects PE decision.

Within our model family. Across the four scales (Baseline AR 100K, QE+APE at 100K, 1M, and

1.84M), Δ BLEU improves with scale on six of the eight pairs. On simplified chinese, japanese, korean, german, and polish, the model trained on 1.84M segments is the best in its column and turns the negative Δ BLEU values of the smaller scales into positive ones (ger: $-15.37 \rightarrow +6.17$; pol: $-12.57 \rightarrow +0.69$; kor: $+1.23 \rightarrow +16.72$). On spa and fre the trend is the same direction (less negative with scale) but Δ BLEU remains negative even at 1.84M (-15.64 and -12.11). For fr_ca the 1M model is best ($+8.09$ BLEU); the 1.84M model is lower ($+1.49$), consistent with the reduced fr_ca share in the rebalanced training set.

Against commercial LLMs. GPT-5.4 and Gemini 3.1 Pro produce negative Δ BLEU on six of the eight pairs, with the largest drops on fr_ca, fre, and spa (Gemini: -20.27 , -21.50 , -23.10). The 1.84M model is positive or close to zero on six of eight pairs and is the best column on five (ger, kor, pol, jap, chi_s tied). The persistent negative Δ BLEU on spa and fre is shared with both commercial models, indicating that the difficulty is not specific to model capacity. We discuss this further in Section 5.5.

5.5 Over-Editing

Two language pairs, eng→spa and eng→fre, show negative Δ BLEU and Δ COMET across all model configurations, including the 1.84M QE+APE model and both commercial LLMs. For these pairs, the original in-house MT output is already strong, with BLEU scores of 82.56 and 61.15 and COMET scores of 1.046 and 0.869, respectively, as shown in Table 2. When the model edits translations that are already close to the reference, even small surface-level changes can reduce BLEU or COMET.

The No-PE subset accuracy in Table 5 further supports this interpretation. The 1.84M model achieves 77.7% accuracy on No-PE examples, meaning that it still edits a non-trivial fraction of translations that should have been left unchanged. This behavior reflects the central challenge of QE-informed post-editing: the model must not only generate good corrections, but also learn when no correction is needed.

The same pattern appears in the commercial LLM baselines, which often produce larger negative Δ BLEU values on high-quality language pairs. Compared with these baselines, our QE+APE models generally reduce the severity of over-editing, but fur-

ther calibration of the PE decision threshold remains an important direction for future work.

Model	QE r	QE MAE	PE Acc
GPT-5.4	0.236	0.165	0.658
Gemini 3.1 Pro	0.310	0.151	0.613
QE+APE (1M) [Ours]	0.863	0.059	0.877
QE+APE (1.84M) [Ours]	0.907	0.054	0.884

Table 8: Global comparison between commercial LLMs and our proposed joint QE+APE models (1M and 1.84M samples).

6 Limitations

While the proposed unified QE+APE architecture demonstrates strong performance, we acknowledge certain limitations in our current experimental setup.

First, our evaluation compares a model fine-tuned on in-domain data against commercial LLMs (GPT-5.4 and Gemini 3.1 Pro) operating in a zero-shot setting. Because the available general purpose models did not have access to the specific domain distribution of our training data, the comparison is inherently asymmetric.

Second, due to resource constraints, our autoregressive baseline was trained only at the 100K scale. While this serves as a proof of concept that explicitly modeling the regression head improves performance over standard autoregressive decoding.

Third, the training targets and test references are derived directly from corrections made by professional linguists. This ensures our Normalized Edit Distance targets accurately represent real world human PE effort. However, evaluation of our generated outputs relies on standard reference based automated metrics (BLEU and COMET).

Finally, regarding reproducibility, the datasets used in this study consist of proprietary in-house data and thus cannot be publicly released. However, our architectural setup and prompt designs are fully detailed to enable replication on public datasets.

7 Future Work

We identify two directions for future work First, re-evaluate available commercial LLMs with few-shot prompting would narrow the data gap without changing their architecture, so any remaining performance difference can be attributed to the unified

QE+APE approach rather than to domain familiarity alone. Second, Extending the head to predict structured like MQM style annotations, error spans, categories, and severities jointly with the correction would scale from a scalar acceptability judgment toward interpretable, error-localized feedback that is more directly useful to linguists.

8 Conclusion

We have shown that jointly fine-tuning a small, specialized language model is an effective approach for translation QE and APE. By integrating quality estimation and post-editing within a single architecture, the model can use a shared representation of translation quality to guide both score prediction and correction generation. This enables more controlled post-editing than a pipeline or zero-shot prompting setup.

Our results show that task-specific joint training is more important than model scale for this setting. The Qwen-based QE+APE models substantially outperform the commercial LLM baselines in prediction accuracy: the 1M and 1.84M models achieve PE decision accuracies of 87.7% and 88.4%, respectively, compared with 65.8% for GPT-5.4 and 61.3% for Gemini 3.1 Pro. The 1.84M model also reaches a much stronger QE correlation, with Pearson’s $r = 0.907$, showing that the proposed regression-head design produces more reliable quality estimates than zero-shot general-purpose LLMs.

Overall, the proposed QE+APE framework reduces the over-editing commonly observed in non-specialized models and provides a more reliable foundation for automatic post-editing in localization workflows. Future work should focus on improving calibration and reducing unnecessary edits, especially for language pairs where the original MT output is already strong.

References

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Chen, P. and Others (2023). Iterative translation refinement with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT)*.
- Feng, Y. and Others (2024). TEaR: Improving LLM-based machine translation with self-refinement. In *Findings of the Association for Computational Linguistics: NAACL 2024*.
- Guerreiro, N. M., Rei, R., and Others (2023). xCOMET: Transparent machine translation evaluation through fine-grained error detection. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*.
- Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Chi, E. H. (2024). Large language models cannot self-correct reasoning yet. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Juraska, J. and Others (2023). MetricX-23: The google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*.
- Kendall, A., Gal, Y., and Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.
- Ki, C. and Carpuat, M. (2024). Guiding large language models to post-edit machine translation with error annotations. In *Findings of the Association for Computational Linguistics: NAACL 2024*.
- Kocmi, T., Bawden, R., Bojar, O., Martins, A. F. T., et al. (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 (EAMT)*.
- Kocmi, T. and Federmann, C. (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.
- Koneru, S., Huck, M., Exel, M., and Niehues, J. (2025). Quality-aware decoding: Unifying quality estimation and decoding. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 33–46.

- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. (2025). Qwen2.5 technical report.
- Ranasinghe, T., Orasan, C., and Mitkov, R. (2020). Transquest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th international conference on computational linguistics*, pages 5070–5081.
- Raunak, V. and Others (2023). Leveraging large language models for automated machine translation post-editing. *arXiv preprint arXiv:2305.xxxx*.
- Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020a). Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)*, pages 2685–2702.
- Rei, R., Stewart, C., Farinha, C., and Lavie, A. (2020b). Unbabel’s participation in the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.
- Zerva, C., Blain, F., De Souza, J. G., Kanojia, D., Deoghare, S., Guerreiro, N. M., Attanasio, G., Rei, R., Orasan, C., Negri, M., et al. (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.
- Zhang, B., Haddow, B., and Birch, A. (2023). Prompting large language model for machine translation: A case study. In *International conference on machine learning*, pages 41092–41110. PMLR.

BRIDGE-MT: A Benchmark for Role Interactions and Dependencies in Machine Translation Gender Evaluation

Neha Gajakos

South East Technological University, Carlow, Ireland

C00315510@setu.ie

Christopher Staff

South East Technological University, Carlow, Ireland

christopher.staff@setu.ie

Brenda Murphy

South East Technological University, Carlow, Ireland

brenda.murphy@setu.ie

John D. Kelleher

ADAPT Centre, Trinity College Dublin, Dublin, Ireland

john.kelleher@adaptcentre.ie

Rejwanul Haque

South East Technological University, Carlow, Ireland

rejwanul.haque@setu.ie

Abstract

This paper investigates gender behavior in Hindi–English machine translation (MT) within multi-entity settings, especially when two occupational roles appear within the same sentence. Existing benchmarks often focus on single-referenced entities, leaving cross-role dependencies largely unexplored. We define a taxonomy of thirteen role-gender configurations covering masculine (*m*), feminine (*f*), and neutral (*n*) assignments and introduce BRIDGE-MT, a manually created dataset of 351 Hindi–English sentence pairs (594 role-level instances) in order to evaluate dual-role interactions. We evaluated commercial MT systems and multilingual LLMs, and propose neutral-comparison asymmetry metrics and a conditional interaction metric to analyze cross-role dependencies. Our results show that explicitly gendered roles achieve higher *F1* scores than neutral-labelled roles. We also observe a consistent position effect, where Role B (the second role) tends to have lower accuracy and greater gender asymmetry than Role A (the first role) across all evaluated systems. Conditional interaction analysis further indicates that the gender assigned to one role can influence the translation of the other. These findings highlight the importance of evaluating gender behavior in multi-entity settings to better understand interaction-driven asymmetries in MT.

1 Introduction

Gender bias in MT has been widely studied in occupational contexts and pronoun resolution (Stanovsky et al., 2019; Prates et al., 2019; Savoldi et al., 2021), where systems frequently default to masculine forms or exhibit asymmetric behavior across masculine and feminine

entities. This has motivated the development of controlled evaluation benchmarks such as WinoMT (Stanovsky et al., 2019), MuST-SHE (Bentivogli et al., 2020), and MT-GenEval (Currey et al., 2022) for analyzing gender realization in MT systems.

Existing benchmarks, such as WinoMT

(Stanovsky et al., 2019), are primarily designed for single-entity evaluation and lack the framework for a systematic examination of interactions between multiple occupational roles in a single sentence. In these multi-entity settings, MT systems face the added complexity of maintaining grammatical coherence across several roles simultaneously. This raises a critical question regarding cross-role dependency: to what extent does the gender realization of one role influence the translation of other roles within the same sentence? Although such benchmarks may contain more than one occupational entity, they typically evaluate only the role referenced by the gendered pronoun, without systematically varying and evaluating the gender assignment of the co-occurring role. For example, in a sentence containing a masculine scientist and a feminine journalist, evaluating both roles allows examination of whether the gender assigned to one role affects the translation of the other.

To systematically investigate these interactions, we define a taxonomy of thirteen role-gender configurations covering m , f , and n assignments across two occupations, denoted as Role A and Role B. Based on this taxonomy, we present **BRIDGE-MT** (Benchmarking Role Interactions and Dependencies in Gender Evaluation for MT),¹ a manually created dataset of 351 Hindi-English sentence pairs designed for structured evaluation of role co-occurrence. Our contributions are as follows: (i) taxonomy and dataset: we define a comprehensive taxonomy and introduce BRIDGE-MT, which treats multiple roles as independent evaluation units, (ii) novel metrics: standard ΔG (Delta Gender) is defined as the difference between $F1$ scores for m and f instances (Stanovsky et al., 2019). We propose neutral-comparison asymmetry metrics that extend the ΔG formulation to contrast gendered and neutral configurations, alongside a conditional interaction metric (ΔI) to quantify cross-role dependencies. and (iii) empirical analysis and findings: we evaluated a range of commercial NMT systems and multilingual LLMs. Our results show that explicitly gendered roles are generally translated more reli-

ably than neutral-labelled roles. We also reveal a consistent position effect, where Role B exhibits lower accuracy and higher asymmetry than Role A. Furthermore, we provide evidence of directional influence, where the gender of one role can significantly impact the translation of the other; we observe that the extent of this interaction varies significantly across different MT systems.

2 Related Work

The evaluation of gender bias in MT has evolved from identifying simple stereotyping to analyzing complex coreference resolution in structured benchmarks. Foremost among these frameworks is WinoMT, which was derived from the Winogender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018) coreference test sets. These benchmarks enable researchers to evaluate whether MT systems preserve gender information for occupational entities under controlled coreference ambiguity. Gender asymmetry is typically quantified using ΔG , defined as the difference between $F1$ scores for m and f instances (Stanovsky et al., 2019). This metric enables a structured comparison of system performance across gender classes.

Saunders et al. (2020) studied gender coreference behavior in MT and showed that gender signals can be over-generalized across entities within the same sentence. Their work focused on controlling gender realization during translation through word-level gender tags and tagged adaptation data. In contrast, our work focuses on evaluating gender behavior under co-occurring occupational roles.

Since WinoMT was originally designed with English as the source language, subsequent work adapted it for other languages. Notably, Singh (2023) introduced WinoMT-Hindi² to evaluate gender behavior in Hindi-English translation. However, while WinoMT-Hindi facilitates evaluation in a morphologically gendered language, its design does not systematically account for the interactions between co-occurring roles. Our work addresses this limitation by introducing a taxonomy-driven benchmark designed for the

¹<https://github.com/BridgeMT/BridgeMT>

²<https://github.com/iampushpdeep/Gender-Bias-Hi-En-Eval>

structured analysis of dual-role interactions and positional dependencies in Hindi–English translation (cf. Section 3.1).

3 The BRIDGE-MT Benchmark

3.1 Limitations of WinoMT-Hindi

Despite its utility in marking source-side gender, the structural design of Singh’s adaptation restricts the evaluation of multi-entity settings in several ways. First, although each sentence in WinoMT-Hindi contains two occupational entities, the English translation typically provides an explicit gender cue for only one of the roles, for e.g.:

Hindi: मजदूर होटल की रिसेप्शनिस्ट से अधिक खाया क्योंकि वह आहार पर है।

Transliteration: Mazdoor hotel ki receptionist se adhik khaya kyunki vah aahaar par hai.

English: The laborer eats more than the receptionist as he is on a diet.

In such cases, while both occupations “मजदूर” (*mazdoor*, ‘laborer’), “रिसेप्शनिस्ट” (*receptionist*, ‘receptionist’) are present, the gendered pronoun in the English translation is associated with only one entity “मजदूर” (*laborer*). As a result, the evaluation focuses primarily on the referenced role “मजदूर” (*laborer*). Other possible configurations of gender assignments across the two occupations are not systematically examined, leaving potential inter-role dependencies unexplored. Second, certain sentences exhibit morphological inconsistencies in gender marking, e.g.:

Hindi: निर्माण मजदूर ने स्कूल की शिक्षक को घूरा क्योंकि उसने निर्माण क्षेत्र में प्रवेश किया।

Transliteration: Nirmaan mazdoor ne school ki shikshak ko ghoora kyunki usne nirmaan kshetra mein pravesch kiya.

English: The construction worker stared at the teacher because she entered the construction zone.

Here, the feminine form “शिक्षिका” (*shikshika*, ‘teacher.f’) would be expected instead of the masculine form “शिक्षक” (*shikshak*, ‘teacher.m’) when referring to a feminine teacher. Such examples highlight the importance of careful morphological validation when constructing gender-aware evaluation datasets. Third, the occupation stereotype groupings in

WinoMT-Hindi are derived from the original English occupation lists in WinoBias. These are based on U.S. Department of Labor statistics, as discussed in Singh (2023). While this provides a structured basis for stereotype categorization, the occupational distributions reflect a U.S.-specific context. Hindi is predominantly spoken in the Indian subcontinent (Kunchukuttan et al., 2018), where certain roles included in the original list, such as “sheriff” or “chiefs” are not widely used or do not directly correspond to culturally prevalent occupational categories. Regional differences in institutional structures and labor distributions may therefore affect how these occupations are interpreted and translated in Hindi.

These observations do not diminish the contribution of WinoMT-Hindi. Instead, they identify opportunities for a more systematic evaluation framework that covers all combinations of gender assignments across co-occurring roles, while ensuring careful lexical and morphological validation. The taxonomy-driven evaluation dataset proposed in this work is designed to address these requirements.

3.2 Taxonomy

We categorize dual-role gender assignments into thirteen distinct configurations, denoted by the grammatical markers assigned to Role A and Role B. This taxonomy provides an exhaustive framework for organizing all possible permutations of m , f , and n labels within a single sentence. In this context, n refers to cases where the source sentence does not provide an explicit gender marker for that role. It represents the absence of gender specification rather than a gender category, and does not capture non-binary or gender-diverse identities. By combining these markers across the two roles, we derive the thirteen configurations presented in Table 1.

ID	Configuration
1	A is masculine
2	A is feminine
3	B is masculine
4	B is feminine
5	A is masculine and B is neutral
6	A is feminine and B is neutral
7	A is neutral and B is masculine
8	A is neutral and B is feminine
9	Both A and B are masculine
10	A is masculine and B is feminine
11	A is feminine and B is masculine
12	Both A and B are feminine
13	Both A and B are neutral

Table 1: Gender-role configurations based on the proposed taxonomy.

This structured approach allows us to move beyond individual role accuracy and quantify how gender assignments for one entity influence the translation of the other. For example, we can examine whether the translation of Role A differs when Role B is *m* versus when Role B is *f* or *n*. Furthermore, symmetric configurations such as “A is *m* and B is *f*” and “A is *f* and B is *m*” allow for a balanced comparison of cross-role behavior and potential asymmetries.

Unlike benchmarks that focus on a single referenced entity per sentence, our proposed taxonomy treats both roles as independent units of evaluation. This design allows both roles to be evaluated independently and jointly, providing the basis for the conditional interaction metric introduced in Section 4.

3.3 Dataset

Based on the taxonomy described in Section 3.2, we constructed a Hindi-English evaluation dataset, BRIDGE-MT, that instantiates the defined role-gender configurations. The objective is to represent controlled gender assignments across co-occurring occupational roles within grammatically natural Hindi sentences.

We selected 27 pairs of occupational roles, denoted as Role A and Role B. These occupations are chosen because they are commonly used in Hindi-language contexts, have clear lexical realizations in both Hindi and English, and are contextually appropriate within the Indian

subcontinent where Hindi is predominantly spoken.

We specifically selected only those occupations that are expressed in gender-neutral forms in Hindi. For example, the role किसान (kisan, ‘farmer’) can refer to both *m* and *f* individuals. By using neutral job titles such as किसान, we ensure that gender is marked through pronouns and verb agreement rather than the occupation name itself. This design choice enabled us to isolate the influence of grammatical gender cues and to measure how accurately an MT model handles them at the sentence level.

For each occupational pairing, sentences are manually constructed to reflect the intended configuration using natural sentence structures. For illustration,

- **A single-role configuration (Role A: *m*)** is realized as:

Hindi: लाइब्रेरियन ने बताया कि वह सूची जाँच करेगा।

Transliteration: Librarian ne bataya ki vah suchi jaanch karega.

English: The librarian said that he would check the list.

Here, the occupation “लाइब्रेरियन” (*librarian*, ‘*librarian*’) is marked with *m* agreement through the verb form “करेगा” (*karega*, ‘*karega*’).

- **A two-role configuration (Role A: *m*, Role B: *f*)** is realized as:

Hindi: वैज्ञानिक ने बताया कि वह उपकरण जाँच करेगा जबकि पत्रकार ने कहा कि वह विश्लेषण लिखेगी।

Transliteration: Vaigyanik ne bataya ki vah upakaran jaanch karega, jabki patrakar ne kaha ki vah vishleshan likhegi.

English: The scientist said he would check the equipment, while the journalist said she would write the analysis.

In this example, gender is encoded separately for each role through verb agreement: *m* through “करेगा” (*karega*, ‘*karega*’) for “वैज्ञानिक” (*vaigyanik*, ‘*scientist*’) and *f* through “लिखेगी” (*likhegi*, ‘*write*’) for “पत्रकार” (*patrakar*, ‘*journalist*’).

- **A mixed configuration (Role A: *m*, Role B: *n*)** is realized as:

Hindi: न्यायाधीश ने कहा कि वह निर्णय कल सुनाएगा, तभी वकील गवाही के कागज़ जुटाने में व्यस्त था।

Transliteration: Nyayadhish ne kaha ki vah nirnay kal sunayega, tabhi vakil gavahi ke kagaz jutane mein vyast tha.

English: The judge said he would deliver the verdict tomorrow, while the lawyer was busy collecting the evidence papers.

In this example, Role A “न्यायाधीश” (*nyayadhish*, ‘judge’) is marked with *m* agreement through the verb form “सुनाएगा” (*sunayega*, ‘would deliver’), while Role B “वकील” (*vakil*, ‘lawyer’) appears without an explicit gender-specific pronoun. Therefore, Role B is treated as *n* within the taxonomy because it does not have a clear gender marking.

The dataset consists of 351 Hindi-English sentence pairs, created from 27 role pairs across 13 configurations. It is designed as a controlled evaluation benchmark rather than a large-scale corpus, with the aim of analyzing interaction-driven gender behavior. Evaluation is conducted at the role level, meaning that sentences containing two occupational roles provide two independent gender predictions. This results in 594 role-level evaluation instances to compare across *m*, *f*, and *n* assignments.

A randomly sampled subset comprising 10% of the source Hindi sentences was independently translated by a second bilingual translator. These translations were then compared with the original references to verify agreement and gender consistency. No discrepancies were observed in the sampled subset.

4 Measuring Interactions and Dependencies

Existing benchmarks such as WinoMT and WinoMT-Hindi evaluate gender behavior primarily through aggregate metrics such as accuracy and ΔG . While this provides insight into overall gender asymmetry, it does not capture dependencies between co-occurring roles within the same sentence.

Since our dataset contains sentences with two occupational roles, it enables conditional analysis. In such settings, the translation behavior for Role A may depend on the gender assignment of Role B, and vice versa. To examine this phenomenon, we propose a conditional interaction metric, denoted as ΔI .

However, interaction can only be meaningfully measured when both roles are explicitly gendered. If one role is neutral or absent, cross-role dependency cannot be evaluated. For this reason, the interaction metric is computed only on the subset of sentences where both Role A and Role B are assigned explicit *m* or *f* gender labels.

4.1 Role-Level Evaluation

In our dataset, the intended gender of each role is predefined by the taxonomy configuration (cf. Section 3.2). We refer to this predefined assignment as the gold gender of the role.

Evaluation is conducted at the role level. For each role $r \in \{A, B\}$, we compute *F1* scores separately for *m* and *f* predictions by comparing system outputs against the gold gender label. Following prior work, overall gender asymmetry for a role is quantified using (1):

$$\Delta G_r^{m-f} = F1_r(m) - F1_r(f) \quad (1)$$

While evaluating MT systems, the absolute value of ΔG_r^{m-f} should ideally be close to zero, indicating balanced performance across *m* and *f* instances. In addition to the standard measure, ΔG_r^{m-f} , we introduce two neutral-comparison asymmetry measures that incorporate neutral-labelled configurations, as in (2) and (3):

$$\Delta G_r^{m-n} = F1_r(m) - F1_r(n) \quad (2)$$

$$\Delta G_r^{f-n} = F1_r(f) - F1_r(n) \quad (3)$$

These measures extend the standard ΔG_r^{m-f} formulation by replacing one gender class with *n* instances. They allow us to quantify how system performance on explicitly gendered roles differs from that in *n* configurations. To the best of our knowledge, neutral-comparison asymmetry measures have not yet been explicitly studied in MT evaluation. While these additional comparisons provide further insight into role-level behavior, they are not used in the computation of the proposed conditional interaction metric.

4.2 Conditional Interaction Definition

To measure interaction, we condition the bias of one role on the gold gender of the other role, using only the fully gendered configurations: (i) both A and B are masculine, (ii) A is masculine

and B is feminine, (iii) A is feminine and B is masculine, and (iv) both A and B are feminine.

For Role B, we compute conditional bias values using (4) and (5):

$$\Delta G_{B|A=m} = F1_{(B=m|A=m)} - F1_{(B=f|A=m)} \quad (4)$$

$$\Delta G_{B|A=f} = F1_{(B=m|A=f)} - F1_{(B=f|A=f)} \quad (5)$$

Similarly, for Role A, we compute $\Delta G_{A|B=m}$ and $\Delta G_{A|B=f}$.

Then the interaction terms are computed using (6) and (7):

$$\Delta I_B = \Delta G_{B|A=m} - \Delta G_{B|A=f} \quad (6)$$

$$\Delta I_A = \Delta G_{A|B=m} - \Delta G_{A|B=f} \quad (7)$$

These quantities effectively measure how the bias associated with one role influences the gender of the co-occurring role. Our proposed interaction metric is interpreted as follows: (i) if $\Delta I_A = 0$ and $\Delta I_B \neq 0$, Role A influences Role B, (ii) if $\Delta I_A \neq 0$ and $\Delta I_B = 0$, Role B influences Role A, (iii) if both are non-zero, mutual interference is present, and (iv) if both are zero, no measurable interaction is observed. Unlike ΔG , which measures overall asymmetry for a single role, ΔI captures cross-role dependency under controlled dual-gender configurations.

5 Experimental Setups

To evaluate gender behavior within our proposed taxonomy, we examined a range of Hindi-English MT systems representing both commercial production environments and academic research. On the commercial side, we evaluated Google Translate³ and Amazon Translate⁴, which are production-level NMT systems available via public interfaces and APIs. We also evaluated IndicTrans2,⁵ (Gala et al., 2023) an open-source multilingual MT system. In addition, we evaluated four

instruction-tuned LLMs capable of performing translation via prompting: Llama 3.1 8B Instruct⁶ (Grattafiori et al., 2024), Qwen3-4B⁷ (Yang et al., 2025), TranslateGemma-12B-IT⁸ (Finkelstein et al., 2026), and Gemma-3-4B-IT⁹ (Team et al., 2025). This selection allows us to compare traditional encoder-decoder MT systems and LLMs under the same controlled evaluation setting.

Each system translates all 351 Hindi sentences into English. Each sentence is translated independently, and the outputs are used directly for evaluation without post-editing or manual correction. For LLMs, translation is performed in a controlled setting to minimize variations. Models are prompted to produce only the English translation of the Hindi input, without additional instructions or paraphrasing. Sampling-based decoding methods are disabled, and decoding parameters are set to ensure consistent outputs. As a result, each model generates a single translation per sentence. This controlled setup reduces decoding-related variance, making model differences easier to interpret.

5.1 Gender Extraction and Role-Level Scoring with Skipped Cases

Evaluation follows the role-level framework defined in Section 4.1. For each role, the gold gender is determined directly from the taxonomy, while the predicted gender is extracted from the translated English output, and identified using explicit markers (e.g., he, him, and his) for m and (e.g., she, her, and hers) for f in the English translations. If no clear marker is present in configurations where gender is specified, the instance is treated as failing to preserve the intended gender. This includes cases where the system produces gender-neutral paraphrases that omit pronouns entirely.

Note that neutral-labelled configurations were also included in the role-level evaluation using the metrics ΔG_r^{m-n} and ΔG_r^{f-n} defined

³<https://translate.google.com>

⁴<https://aws.amazon.com/translate>

⁵<https://github.com/AI4Bharat/IndicTrans2>

⁶<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

⁷<https://huggingface.co/Qwen/Qwen3-4B>

⁸<https://huggingface.co/google/translategemma-12b-it>

⁹<https://huggingface.co/google/gemma-3-4b-it>

in Section 4.1. Additionally, if a system’s output does not contain a lexical form that corresponds to the gold occupation, the instance is marked as a skipped case and excluded from scoring. These instances typically occur when the target occupation undergoes mistranslation, substitution, or lexical alteration, resulting in a loss of semantic correspondence with the intended role. For instance, consider the following translation:

Source sentence (Hindi): शेफ़ ने संकेत किया कि वह स्वाद-परख पूरी करेगा, तदनंतर सचिव ने बताया कि वह विवरण-पत्र को अंतिम रूप देगी।

Transliteration: Chef ne sanket kiya ki vah swaad-parakh poori karega, tadanantar sachiv ne bataya ki vah vivaran-patra ko antim roop degi.

English (Gold): The chef indicated that he would complete the tasting, after which the secretary said she would finalize the detail sheet.

Qwen3-4B Output: Sheikh indicated that he would complete the taste test, and then the secretary said that she would finalize the details letter.

In this case, Role A (“chef”) is mistranslated as “Sheikh”. Since the translated word does not correspond to the intended occupation, the gender associated with that role cannot be reliably evaluated. Therefore, Role A is treated as a skipped case and excluded from scoring.

For role-level metrics, valid roles are evaluated independently even if the other role in the sentence is skipped. However, for the conditional interaction metric, both roles must be correctly identified. If either role is skipped, the sentence is excluded from interaction analysis.

5.2 Evaluation Metrics

We measured role-level accuracy and gender asymmetry ΔG^{m-f} , following the standard benchmark evaluation framework. We computed ΔG^{m-n} and ΔG^{f-n} to analyze performance differences between gendered and neutral configurations. The interaction metric (ΔI), defined in Section 4, was computed on the fully gendered dual-role subset. To provide an overall measure of translation quality, we additionally report BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), and COMET (Rei et al., 2020) scores for all evaluated systems.

System	chrF++	BLEU	COMET
Google Translate	63.61	41.95	0.8210
Amazon Translate	59.06	34.42	0.8125
IndicTrans2	60.13	37.16	0.8118
Llama 3.1 8B Inst.	56.67	29.54	0.8020
Qwen3-4B	55.02	29.82	0.8028
TranslateGemma-12B	57.13	29.80	0.8153
Gemma-3-4B	57.91	32.15	0.8099

Table 2: Performance of the MT systems on the BRIDGE-MT dataset.

6 Results and Discussions

We report results for all evaluated systems on the BRIDGE-MT dataset, using evaluation metrics discussed in Section 5.2. Table 2 presents translation quality results and serves as a general performance baseline across systems. While all systems maintain relatively high translation quality, there is notable variance in scores between the top-performing commercial engines and LLMs. Tables 3, 4, and 5 summarise overall role-level performance (combined and role-wise), while Table 6 presents conditional interaction results on the fully gendered subset.

6.1 Overall Role-Level Performance

Table 3 presents overall results across both roles. Among the evaluated MT systems, Google Translate achieves the highest overall accuracy (77.21%), followed by IndicTrans2 (74.16%) and Amazon Translate (70.82%). Google Translate and IndicTrans2 exhibit nearly balanced performance across m and f instances, with ΔG^{m-f} values close to zero (-0.0005 and 0.0022 , respectively). In contrast, Llama 3.1 8B Instruct shows a large positive ΔG^{m-f} (0.2336), reflecting substantially higher $F1$ for m instances than f instances. Gemma-3-4B-IT also exhibits noticeable asymmetry ($\Delta G^{m-f} = 0.1293$). Qwen3-4B and TranslateGemma-12B-IT show smaller but still positive ΔG^{m-f} values.

Across all systems, performance on neutral instances is lower than on explicitly gendered cases. This suggests that the absence of explicit gender cues introduces additional difficulty in translation. Skipped cases remain relatively low for most systems, with the exception of Qwen3-4B (71 skipped cases) and Llama 3.1 8B Instruct

System	Acc (%)	$F1(m)$	$F1(f)$	$F1(n)$	ΔG^{m-f}	ΔG^{m-n}	ΔG^{f-n}	Skipped
Google Translate	77.21	0.8438	0.8443	0.4486	-0.0005	0.3952	0.3958	6
Amazon Translate	70.82	0.7717	0.7524	0.4579	0.0194	0.3138	0.2944	1
IndicTrans2	74.16	0.8048	0.8026	0.4706	0.0022	0.3342	0.3320	2
Llama 3.1 8B Instruct	51.05	0.6028	0.3691	0.4278	0.2336	0.1749	-0.0587	26
Qwen3-4B	69.79	0.7561	0.7362	0.4848	0.0199	0.2713	0.2514	71
TranslateGemma-12B-IT	68.13	0.7524	0.7378	0.4298	0.0146	0.3227	0.3081	1
Gemma-3-4B-IT	54.82	0.6293	0.5000	0.4240	0.1293	0.2053	0.0760	3

Table 3: Overall role-level results across both roles combined (594 instances). Accuracy, $F1$ scores for masculine (m), feminine (f), and neutral (n) instances, overall gender asymmetry (ΔG), neutral-comparison asymmetry measures (ΔG^{m-n} , ΔG^{f-n}), and number of skipped cases are reported.

System	Acc (%)	$F1(m)$	$F1(f)$	$F1(n)$	ΔG^{m-f}	ΔG^{m-n}	ΔG^{f-n}	Skipped
Google Translate	80.28	0.8770	0.8710	0.4285	0.005	0.4485	0.4425	1
Amazon Translate	77.49	0.8307	0.8358	0.4587	-0.0051	0.3721	0.3771	0
IndicTrans2	81.66	0.8733	0.8888	0.4727	-0.0155	0.4006	0.4161	2
Llama 3.1 8B Instruct	57.49	0.6560	0.4888	0.4242	0.1672	0.2318	0.0646	24
Qwen3-4B	72.96	0.7875	0.7665	0.4800	0.0209	0.3075	0.2865	44
TranslateGemma-12B-IT	72.57	0.7973	0.7777	0.4354	0.0196	0.3619	0.3422	1
Gemma-3-4B-IT	56.73	0.6611	0.5192	0.3970	0.1417	0.2639	0.1222	2

Table 4: Role-wise evaluation: Role A only

System	Acc (%)	$F1(m)$	$F1(f)$	$F1(n)$	ΔG^{m-f}	ΔG^{m-n}	ΔG^{f-n}	Skipped
Google Translate	72.69	0.7917	0.8022	0.4706	-0.0105	0.3210	0.3316	5
Amazon Translate	61.16	0.6872	0.6053	0.4571	0.0820	0.2301	0.1481	1
IndicTrans2	63.37	0.7029	0.6589	0.4684	0.0440	0.2345	0.1904	0
Llama 3.1 8B Instruct	42.32	0.5267	0.1864	0.4313	0.3402	0.0953	-0.2449	2
Qwen3-4B	65.28	0.7106	0.6861	0.4897	0.0245	0.2208	0.1963	27
TranslateGemma-12B-IT	61.72	0.6859	0.6708	0.4237	0.0151	0.2622	0.2470	0
Gemma-3-4B-IT	52.07	0.5812	0.4706	0.4561	0.1106	0.1251	0.0144	1

Table 5: Role-wise evaluation: Role B only

System	$\Delta G_{B A=m}$	$\Delta G_{B A=f}$	ΔI_B	$\Delta G_{A B=m}$	$\Delta G_{A B=f}$	ΔI_A
Google Translate	0.0333	-0.0014	0.0348	0.0043	0.0000	0.0043
Amazon Translate	0.6036	0.0168	0.5868	0.0342	0.0231	0.0111
IndicTrans2	0.3105	-0.0387	0.3492	-0.0077	-0.0014	-0.0063
Llama 3.1 8B Instruct	0.2778	0.4703	-0.1926	0.1065	0.0768	0.0296
Qwen3-4B	0.0439	0.0873	-0.0435	0.0108	0.0366	-0.0258
TranslateGemma-12B-IT	0.1357	0.0015	0.1342	0.0687	-0.0222	0.0909
Gemma-3-4B-IT	0.2456	0.1392	0.1063	0.2193	0.1576	0.0618

Table 6: Conditional Interaction Results (Fully Gendered Subset)

(26 skipped cases).

6.2 Role-Wise Analysis

Tables 4 and 5 present results for Role A and Role B respectively. This breakdown allows us to examine whether systems behave differently depending on entity position. For Role A (Ta-

ble 4), MT systems perform strongly. IndicTrans2 achieves the highest accuracy (81.66%), followed closely by Google Translate (80.28%) and Amazon Translate (77.49%). ΔG^{m-f} values for these systems are small in magnitude, indicating balanced $m-f$ behavior for the first role. As for Role B (Table 5), performance de-

creases across all systems. Google Translate accuracy drops from 80.28% (Role A) to 72.69% (Role B). Amazon Translate shows a substantial decrease (77.49% to 61.16%), and IndicTrans2 similarly declines (81.66% to 63.37%).

Overall, the role-wise breakdown reveals a consistent position effect across systems. Role B not only achieves lower accuracy than Role A but also, in several cases, exhibits larger gender asymmetry. This observation motivates the need for conditional interaction analysis.

6.3 Conditional Interaction Results

Table 6 presents conditional interaction scores computed on the fully gendered subset. For Google Translate, both ΔI_B (0.0348) and ΔI_A (0.0043) are close to zero. This indicates relatively stable behavior, as the gender assigned to one role has minimal effect on the gender realization of the other role. Amazon Translate, however, shows a very large ΔI_B value (0.5868), while ΔI_A remains small (0.0111). This suggests strong directional interaction, with the gender of Role A substantially influencing the gender realization of Role B, but not vice versa. IndicTrans2 also shows noticeable interaction for Role B ($\Delta I_B = 0.3492$), though smaller than Amazon Translate. Interaction for Role A remains negligible (-0.0063). Among LLM-based systems, interaction patterns vary. Some models show moderate interaction in both directions, while others show limited dependency between roles.

A consistent pattern across several systems is that $|\Delta I_B|$ tends to be larger than $|\Delta I_A|$. This aligns with the earlier role-wise findings, which indicated that Role B showed greater instability and lower accuracy. Overall, the second role in the sentence appears more sensitive to the gender of the first role.

7 Conclusions

In this paper, we investigated gender bias in Hindi–English MT within multi-entity contexts. By introducing a structured taxonomy and the BRIDGE-MT dataset, we moved beyond single-role evaluation to quantify how co-occurring occupational roles influence translation outcomes.

Our empirical analysis showed that MT systems perform more consistently on explicitly

gendered configurations than on neutral ones. We also observed a consistent position effect: Role B typically shows lower accuracy and, in several cases, larger gender imbalance than Role A. Conditional interaction analysis further suggests that cross-role effects vary across the MT systems, with some systems exhibiting stable behavior and others showing strong directional influence, most often from Role A to Role B.

Overall, this work highlights the importance of evaluating gender behavior beyond single-entity settings, where dependencies between roles may affect translation outcomes.

As future work, we plan to explore mitigation strategies for reducing interaction-driven asymmetries and to extend this analysis to additional languages, including other gender-marking language pairs and larger datasets.

References

- Bentivogli, L., Savoldi, B., Negri, M., Di Gangi, M. A., Cattoni, R., and Turchi, M. (2020). Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.
- Currey, A., Nadejde, M., Pappagari, R. R., Mayer, M., Lauly, S., Niu, X., Hsu, B., and Dinu, G. (2022). MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Finkelstein, M., Caswell, I., Domhan, T., Peter, J.-T., Juraska, J., Riley, P., Deutsch, D., Kovacs, G., Dilanni, C., Cherry, C., et al. (2026). Translatagemma technical report. *arXiv preprint arXiv:2601.09012*.
- Gala, J., Chitale, P. A., Ak, R., Gumma, V., Doddapaneni, S., Kumar, A., Nawale, J., Sujatha, A., Puduppully, R., Raghavan, V., et al.

- (2023). Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. (2018). The IIT Bombay English-Hindi parallel corpus. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Popović, M. (2017). chrF++: words helping character n-grams. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., and Kreutzer, J., editors, *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Prates, M. O. R., Avelar, P. H. C., and Lamb, L. (2019). Assessing gender bias in machine translation – a case study with google translate.
- Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). Gender bias in coreference resolution. In Walker, M., Ji, H., and Stent, A., editors, *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.
- Saunders, D., Sallis, R., and Byrne, B. (2020). Neural machine translation doesn’t translate gender coreference right unless you make it. In Costajussà, M. R., Hardmeier, C., Radford, W., and Webster, K., editors, *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.
- Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., and Turchi, M. (2021). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.
- Singh, P. (2023). Gender inflected or bias inflicted: On using grammatical gender cues for bias evaluation in machine translation. In Li, D., Mahendra, R., Tang, Z. P., Jang, H., Murawaki, Y., and Wong, D. F., editors, *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: Student Research Workshop*, pages 17–23, Nusa Dua, Bali. Association for Computational Linguistics.
- Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.
- Team, G., Kamath, A., Ferret, J., et al. (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Yang, A., Li, A., Yang, B., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In Walker, M., Ji, H., and Stent, A., editors, *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.

Improving Term Evaluation in Machine Translation: Variation Matters

Nicolas Dahan

Inria, Paris, France

nicolas.dahan@inria.fr

Ziqian Peng

Sorbonne Université, CNRS, ISIR, Paris, France & Inria, Paris, France

peng@isir.upmc.fr

François Yvon

Sorbonne Université, CNRS, ISIR, Paris, France

yvon@isir.upmc.fr

Rachel Bawden

Inria, Paris, France

rachel.bawden@inria.fr

Abstract

Terminology evaluation in machine translation (MT) usually assumes a single correct target form per source term. However, human translators routinely introduce variation that current metrics penalize as inconsistency. We examine how to account for this variation in document-level MT evaluation of English–French scientific translation, combining glossary-based accuracy, translation consistency, and a new *cross-term variation* (CTV) diagnostic measure that tests whether variation relationships are preserved across languages. Based on analyses of two parallel corpora, translated by four MT systems, we find that (1) MT systems generate less target-side variation than human translators; (2) transfer patterns strongly depend on the variation type; (3) consistency rankings vary with the choice of metric; and (4) constraining MT with a glossary improves accuracy and consistency but degrades CTV by suppressing valid variation. We argue for variation-aware evaluation that conditions consistency penalties on whether target-side variation mirrors source-side variation.

1 Introduction

Evaluating how machine translation (MT) systems handle terminology in scientific and specialized documents requires accounting for a fundamental property of terms in running texts: *terminological variation*. A concept (e.g., *machine translation*) does not always appear under a single form, which we call the preferred term. It may surface as an acronym (*MT*), a reduction (*translation*), a lexical substitution (*automatic translation*), or a morphosyntactic reformulation (e.g., *translation by machines*) (Daille, 2017). We refer to these alternative surface forms as variants. Such variation arises naturally in scientific writing, for example because of stylistic alternation, register adaptation, or avoidance of repetition, and is a well-attested feature of specialized discourse (Fernández-

Silva and Kerremans, 2011; Daille, 2017).

Whether such variation should be considered an error depends on the role the terminological resource is meant to play. When a glossary prescribes a single sanctioned form, for reasons of trademark, safety, or usability, consistency is genuinely required and any deviation is an error, on both the source and the target side. However, glossaries are just as often provided as vocabulary guidance, and terminology standards themselves recognize a scale of acceptability, from preferred to merely admitted forms (ISO, 2019). Unless terminology is enforced to be unambiguous and used consistently, as in controlled-language settings, stylistic variation is expected and even preferred, as in the scientific corpora we study. Yet most evaluation metrics penalize such variation uniformly, ignoring whether it mirrors variation already present in

the source.

Current MT terminology evaluation does not account for variation. *Glossary-based methods* (ibn Alam et al., 2021; Semenov et al., 2023, 2025) match source terms against expected translations from a bilingual glossary, but cannot evaluate non-glossary variants, i.e. a concept’s surface forms that are not listed in the glossary. Such forms naturally arise in running texts, such as reorderings (*quality of translation* instead of *translation quality*), reductions (*processing* for *language processing*), or context-dependent acronyms. Because such forms are produced dynamically as the text unfolds (by insertion, reduction, or new terms being coined from existing ones), they cannot be exhaustively enumerated in advance, and simply enlarging the glossary therefore cannot solve the problem. *Consistency metrics* (Itagaki et al., 2007; Lyu et al., 2021; Semenov and Bojar, 2022) assess whether a source term is uniformly rendered throughout a document, but considers all target-side variation as inconsistency. Neither family of metrics captures whether the *variation relationship* between a preferred term and its variant, for example between an acronym and its expanded form, is preserved in translation; treating the two as separate terms only measures each in isolation.

In this work, we study how to account for terminological variation in document-level MT evaluation for English–French scientific translation. We extend terminology coverage to non-glossary variants, identified in context by a concordancer (Section 3), and examine three complementary evaluation dimensions: (1) term-level accuracy, (2) translation consistency, both established in prior work, and (3) cross-term variation, a new dimension introduced to assess whether variation relationships are preserved in translation (see Table 1). Each dimension has structural coverage gaps tied to variation type, but together they achieve wider coverage. We analyze two parallel English–French scientific corpora from the Natural Language Processing (NLP) domain (Salesky et al., 2023; Peng et al., 2026), each translated by four MT systems, and report the following findings:

- Human translators introduce more target-side lexical variation than MT systems, reflecting standard translation practices that current consistency metrics penalize.
- Variation types are not always preserved in trans-

lation, with some types (e.g., reductions) transferred consistently and others (e.g., acronyms) strongly dependent on the corpus and its translation guidelines.

- Different measures of translation consistency result in divergent system rankings, showing that existing consistency scores are sensitive to metric design.
- Constraining MT to use glossary terms improves accuracy and consistency metrics, but reduces the system’s ability to transfer source-side variations onto the target, exposing a tension between terminological uniformity and valid variation.

Our contribution is primarily methodological: rather than claim that these variation patterns are universal, we provide a framework and metrics for measuring how terminological variation interacts with MT evaluation, and report empirical findings specific to the two English–French NLP corpora we study.

Head (EN)	automatic translation system
Head (FR)	système de traduction automatique
Glossary variants	none listed
Src variant	translation system (<i>reduction</i>)
Expected tgt	système de traduction (<i>reduction, inferred</i>)
MT output	système de traduction (<i>unevaluated by glossary-based methods</i>)

Table 1: Cross-term coherence example: the reduction *translation system* is absent from the glossary, but its French counterpart is inferred from the reduction relationship to the head term.

2 Related Work

Modern MT evaluation increasingly relies on fine-tuned neural metrics such as COMET (Rei et al., 2020) and MetricX (Juraska et al., 2024), which provide strong overall quality estimates, while LLM-based error analysis approaches such as GEMBA-MQM (Kocmi and Federmann, 2023) produce fine-grained Multidimensional Quality Metrics (MQM) annotations that include terminology errors. These methods operate at the segment level, however, and do not model term-level structure: they neither track per-concept cross-document consistency nor capture

variation relationships between preferred terms and variants. This holds most directly for the neural metrics, which score semantic adequacy rather than surface form: they rarely penalize valid variation but, for the same reason, cannot diagnose terminological inconsistency. GEMBA-MQM goes beyond adequacy, targeting terminology through a dedicated MQM category, but its predicted error spans do not align well with human annotations (Lu et al., 2025), and, scoring each segment in isolation, it does not track whether a concept is rendered consistently across a document or whether variation relationships are preserved. Dedicated term-level evaluation protocols have therefore been developed to close these gaps.

A first family of term-level methods relies on a bilingual glossary, measuring how often MT produces the expected target term for each source entry. Matching strategies range from exact surface-form matching (ibn Alam et al., 2021) to partial and lemma-based matching (Alam et al., 2021; Semenov et al., 2025), and have been adopted in shared evaluation campaigns such as the WMT terminology translation tasks (Alam et al., 2021; Semenov et al., 2023, 2025). Document-level extensions further compute per-term ratios of target-to-source occurrences (Semenov et al., 2025). Even with lemmatization, these approaches remain centered on canonical glossary entries and miss unlisted surface forms such as reductions, dynamic acronyms, and lexical variants.

Glossaries are also useful to enforce specific translations at generation time through lexically constrained decoding (Hokamp and Liu, 2017; Post and Vilar, 2018) or constraint-aware training (Dinu et al., 2019), with related lexical constraints used in paraphrase generation (Hu et al., 2019); our glossary-guided condition (Section 5) instead applies such guidance through prompting rather than hard decoding constraints.

When no glossary is available, consistency-based metrics offer an alternative. Semenov and Bojar (2022) compute per-term consistency with respect to a pseudo-reference derived from the first target realization. Concentration-based metrics, such as the Herfindahl–Hirschman Index (Itagaki et al., 2007; Gašpar et al., 2022) and the Lexical Translation Consistency Ratio (LTCR) (Lyu et al., 2021; Wang et al., 2025) quantify how uniformly a source term is translated, by rewarding the reuse of a dominant form. Despite methodological differences, these methods

share a common assumption: optimal scores require that every occurrence of a term should receive the same translation. They therefore penalize acceptable target-side variation, such as the stylistic alternation commonly observed in scientific writing (Vinay and Darbelnet, 1995), where alternating between forms is a deliberate, well-motivated device rather than carelessness (Bowker, 1998; Pecman, 2014), and enforcing a single form risks over-standardization that erases author-intended distinctions (Bowker and Hawkins, 2006).

This assumption sits at odds with a long line of linguistic work showing that terminological variation is a fundamental property of specialized discourse rather than random noise. Daille (2017) provides a comprehensive characterization of denominative variation in domain-specific corpora, distinguishing graphical, morphosyntactic, semantic, and pragmatic variants of a single concept; Cárcamo et al. (2025) refine this typology into graphical, morphosyntactic, reduction, expansion, and lexical variants. Translation studies have also started to examine how variation behaves under translation: Fernández-Silva and Kerremans (2011) show that source-side variation is broadly preserved or accentuated in human translations of Galician scientific texts into English, while Čulo and Nitzke (2016) observe, in English–German post-edited texts, that MT systems produce fewer variant translations than humans, with MT patterns carrying over into post-editing. This reduction is consistent with broader evidence that MT produces lexically less diverse and less rich output than human translators (Vanmassenhove et al., 2019, 2021), and with the notion of *translationese*, of which post-edited MT is an exacerbated case: it is simpler and more normalized than human translation (Toral, 2019). These descriptive findings have not yet been integrated into automatic evaluation: none of these metrics model how variation types affect evaluation coverage, nor whether structural relationships between preferred terms and variants (e.g., acronym–expansion pairs) are preserved in translation.

3 Methodology

We study how MT systems translate terms at the document level, accounting for the range of surface forms a term may assume in context, along three complementary dimensions: (1) how accurately each occurrence is translated based on a reference glossary,

(2) how consistently each surface form (whether a preferred term or a variant) is rendered throughout its document occurrences, and (3) whether variation relationships between terms and variants are preserved in translation.

Our terminological resource is structured as a Simple Knowledge Organization System (SKOS)¹ glossary: each entry pairs a concept with a preferred term (SKOS `prefLabel`) and a list of alternative terms (SKOS `altLabel`); we use *variant* as a synonym for alternative term. We turn term occurrences into their canonical forms by stripping inflection and other non-terminological variation (e.g., by lemmatization and orthographic normalization). Our pipeline takes this glossary as input and processes parallel documents in four steps: (1) term detection, (2) source-side variation labeling, (3) source-target alignment, and (4) target-side labeling. The output is a list of annotated term pairs relating each source occurrence with the corresponding target span.

Type	Description and example
GRAPH	Graphical (acronyms, spelling): <i>intellectual property / IP</i>
MORPH	Morphosyntactic (reordering, inflection): <i>contract of employment / employment contract</i>
RED	Reduction (truncation): <i>European Union acquis / acquis</i>
EXP	Expansion (added material): <i>wear / normal wear and tear</i>
LEX	Lexical (near-synonymy): <i>action for damages / claim for damages</i>
COMBO	Combined (multiple types): <i>coronavirus disease 2019 / COVID-19</i>

Table 2: Typology used to label variation in source- and target term occurrences (Cárcamo et al., 2025).

Term detection. We identify occurrences of known terms using CONCORDANCER,² a trie-based concordancer operating on documents to which linguistic analysis (lemmatization, POS tagging, dependency parsing) is applied with spaCy (Honnibal et al., 2020). Coverage extends beyond exact glossary entries via two complementary mechanisms: (1) *static*

variants, generated offline from each glossary entry by applying a fixed set of transformation rules (synonym substitution via WordNet (Miller, 1995), syntactic restructuring, morphological derivation via MorphyNet (Batsuren et al., 2021), and modifier reduction); (2) *dynamic variants*, detected at run-time from dependency patterns in context. Together these recover a wider range of forms under which each concept appears in the document.

Source-side variation labeling. Each detected occurrence is labeled with its variation type relative to the preferred term, following the typology of Cárcamo et al. (2025) (Table 2; full typology in Appendix A.14). The label provenance depends on whether the matched form is lexicalized in the glossary: for the preferred term and its alternative terms, we assign the type by few-shot prompting ChatGPT (gpt-4.1-mini; prompts in Appendix 9); for non-lexicalized matches (static or dynamic variants), the type is derived directly from the CONCORDANCER formation pattern that triggered the match. When a non-lexicalized match is derived from an alternative term rather than from the preferred term, the final label combines both classifications (e.g., a dynamic acronym of an EXP variant becomes COMBO). To validate this step, one of the authors, a computational linguist and native French speaker fluent in English, manually assigned gold variation labels to 50 sampled pairs; the classifier agrees with these human labels on 84% of types and 77% of subtypes, with near-perfect agreement on single-type variants and systematic under-detection of COMBO (recall 11%; see Appendix B). Since COMBO accounts for fewer than 2% of occurrences, this has little impact on aggregate scores.

In our corpora, exact-form matching accounts for about 62–63% of term occurrences, with lemma matching increasing these numbers to about 82–87%. Additional matches contributed by CONCORDANCER (static and dynamic variants) are dominated by lexical (LEX) and reduction (RED) forms (full breakdown in Appendix A.1).

Alignment. We apply Bertalign (Liu and Zhu, 2022) for sentence-level alignment, and the token-

¹<https://www.w3.org/TR/skos-reference/>

²<https://gitlab.inria.fr/almanach/concordancer>

³We use the HuggingFace models `paraphrase-multilingual-mpnet-base-v2` and `facebook/xlm-roberta-xl` as the embedding models for Bertalign and SimAlign, respectively.

level correspondences are computed with SimAlign (Jalili Sabet et al., 2020).³ For each source term occurrence, the aligned target span is assembled from individual token alignments. When the cosine similarity of the assembled span falls below 0.8, the bilingual sentence pair, the source term, and the SimAlign candidate are submitted to ChatGPT (gpt-4.1-mini; prompt in Appendix 2); the LLM returns the translated form of the source term as it appears in the target sentence (an empty return signals an omission), which replaces the SimAlign span. The 0.8 cutoff is a conservative operating point that reserves the LLM fallback for spans where the embedding aligner is least confident, keeping higher-similarity spans unchanged. To validate this step, the same annotator evaluated 100 term alignments sampled across the nine system outputs; 96% (96/100) were correct, with the four errors limited to span-boundary and lexical-choice mistakes on single occurrences (see Appendix C).

Target-side labeling. Target-side analysis combines two parallel processes that are reconciled afterwards. (1) CONCORDANCER scans the full target document and labels each detected occurrence using the same two-level protocol as the source side (CONCORDANCER + ChatGPT), producing a variation label relative to the expected target preferred term. (2) The alignment step assembles, for each source term occurrence, a target span. We reconcile the two by matching each canonicalized aligned span against CONCORDANCER’s target term list: all aligned spans found by CONCORDANCER inherit its variation label, while aligned spans missed by CONCORDANCER are directly labeled by ChatGPT, given the source term and the aligned span. Each realization is additionally assigned a glossary conformity label: *glossary-conforming*, *variant translation*, or *out-of-glossary* (paraphrase, omission, or error).

4 Evaluation Framework

We analyse source-target associations of terms along three measurements: *term-level accuracy* (Dimension 1, building on glossary-based exact match (ibn Alam et al., 2021)), *translation consistency* (Dimension 2, building on the consistency tradition (Itagaki et al., 2007; Lyu et al., 2021; Semenov and Bojar, 2022)), and *cross-term variation* (Dimension 3, introduced here). These quantities serve a dual purpose. Applied to the human translation references, they de-

scribe terminological behaviour in a corpus; applied to MT outputs against the source and the reference, they evaluate translation quality. The descriptive use establishes that the human reference is not arbitrary; the evaluative use asks whether MT systems resemble human decisions, and whether such similarities are even desirable.

Dimension 1: Term-level accuracy. We compute Exact Match (EM) accuracy: the proportion of occurrences whose translation matches the glossary entry. EM can be applied to any occurrence whose canonical form maps to a glossary entry (preferred term or alternative term).

Dimension 2: Translation consistency. For each canonical surface form (that is, each preferred term or variant treated independently) with at least two occurrences, we collect all its target counterparts in a document and measure the consistency of the resulting translation set. We apply three metrics to compute a translation consistency ratio (TCR); none of them account for whether the target-side variation is justified by source-side differences.

- **TCR_{EMmacro} (TCR-EM):** for each term with $k \geq 2$ occurrences, computes the proportion of occurrences that match the glossary, then averages across terms (term-level).
- **TCR temporal (TCR-T):** for each pair of successive occurrences of the same source canonical form (the 1st and 2nd, then the 2nd and 3rd, etc.), TCR-T checks whether they receive identical translations. The score is the proportion of consecutive pairs that agree.
- **TCR pseudo-ref (TCR-PR):** consistency measured against a pseudo-reference form (the first occurrence). This score does not require a glossary and applies even when a terminological resource is lacking.

Of these, TCR-PR implements the pseudo-reference consistency of Semenov and Bojar (2022), TCR-EM applies glossary-based EM (ibn Alam et al., 2021) at the term level, and TCR-T is a temporal variant we adopt to capture whether translation shifts are gradual or repeated; we retain all three, rather than select one, because their divergence is itself a finding: the choice of consistency definition is not neutral and can change system rankings (Section 6.2). TCR-EM

combines Dimensions 1 and 2: occurrences must agree with each other and with the glossary. This makes it more restrictive than TCR-T and TCR-PR: a term consistently translated with a correct but non-glossary variant translation would score 0 despite being perfectly consistent. TCR-T and TCR-PR are purely consistency-oriented: they measure whether a term is translated uniformly, regardless of which translation is used. Table 20 (Appendix A.8) illustrates how these metrics can diverge on the same data.

Dimension 3: Cross-term variation (CTV). While the previous dimensions assess each preferred term or variant independently (each treated as a separate canonical surface form), this dimension evaluates whether the variation relationships between a preferred term and its variants are transferred to the target. Specifically, we measure *Variation preservation*: when a source preferred term is correctly translated, each GRAPH/EXP/RED/LEX variant should be translated as a corresponding form in the target. We retain these four types because each has a predictable expected target form: a source acronym is typically rendered as a target acronym (GRAPH), a truncated source form as a target truncation (RED), an expanded source form as a target expansion (EXP), and a near-synonymous substitution as a target lexical variant or the target preferred term itself (LEX).⁴

Metric	Definition	Exclusions
1. Term-Level Accuracy		
EM (micro)	<i>Occurrence-level conformity</i>	OOG terms
2. Translation Consistency		
TCR-PR	<i>Match to first occurrence</i>	$k_t < 2$
TCR-T	<i>Consecutive stability</i>	$k_t < 2$
TCR-EM	<i>Avg. conformity per term</i>	$k_t < 2 + \text{OOG}$
3. Cross-Term Variation		
CTV	<i>Preserve variation type</i>	MORPH, COMBO

Table 3: Evaluation dimensions. k_t : number of occurrences of term t . OOG: out-of-glossary.

Table 3 summarizes all metrics, their objectives,

⁴MORPH and COMBO are excluded because their source-to-target mapping is less predictable: morphosyntactic changes depend on target-language grammar, and compound variants involve multiple simultaneous transformations.

⁵Traitement Automatique des Langues Naturelles, <https://www.atala.org/-Conference-TALN-RECITAL>.

⁶Journées d’Etude de la Parole, <https://www.afcp-parole.org/category/animations-scientifiques/les-jep/>.

⁷<https://2022.aclweb.org>.

⁸<https://skosmos.loterre.fr/8LP/fr/>

and exclusion criteria. Importantly, the framework is diagnostic: it characterizes where and how current evaluation clashes with term variation but does not assess translation quality. A term may be consistently translated yet semantically wrong, or inconsistently translated in a contextually appropriate manner.

5 Experimental Setup

We conduct our analysis on two English–French parallel corpora in the NLP domain: ParaNLP and IWSLT2023. Table 4 summarizes their statistics.

ParaNLP The ParaNLP corpus (Peng et al., 2026) consists of 32 English papers published in NLP conferences (2002–2015) paired with comparable French versions published in TALN⁵ or JEP⁶ conferences, resulting in 6,272 parallel sentences. It was constructed via automatic sentence alignment, with Automatic Post-Editing and MT used to handle poorly aligned segments (details in Appendix A.12).

IWSLT2023 The IWSLT2023 corpus comprises the validation and evaluation sets of the IWSLT 2023 translation shared task (Salesky et al., 2023). It contains manually revised transcripts and translations of 10 presentations delivered at the ACL 2022 conference,⁷ aligned at the sentence level. We use the English–French subset.

Terminological resource. Our terminological reference is a manually curated bilingual glossary covering 1,642 NLP concepts with their preferred English and French realizations, compiled from domain-specific resources and validated by experts. It was prepared and published by INIST, the French Institute for Scientific and Technological Information.⁸

Models and Inference. We evaluate four medium-size multilingual LLMs using vLLM (Kwon et al., 2023) (model cards in Appendix A.11): Llama3.1-8B-Instruct (Llama) (Grattafiori et al., 2024), Qwen3-8B (Qwen) (Yang et al., 2025), EuroLLM-9B-Instruct (Euro9) (Martins et al., 2025), and EuroLLM-22B-Instruct (Euro22) (Ramos et al., 2026). We refer to the human reference as “Ref”. We ap-

ply greedy decoding to all systems except Qwen, for which we use the recommended “non-thinking” mode (temperature=0.7, top- p =0.8, top- k =20).

Translation settings. We evaluate MT outputs under two zero-shot conditions: (1) baseline, where models translate (English-to-French) without terminological guidance; and (2) glossary-guided (base+terms), where glossary-preferred terms and their expected translations are injected into the prompt. Comparing the two conditions serves as a diagnostic for whether constrained systems genuinely exploit the glossary information (prompts in Appendix 3).

6 Analysis

6.1 Variation Patterns: Reference and MT

Statistic	ParaNLP	IWSLT2023
<i>Corpus overview</i>		
#docs	32	10
Avg. #sents / doc	196	88
Avg. concepts / doc	50.44	31.00
<i>Occurrence & variation richness</i>		
Avg. occ. / concept	5.55	3.92
Avg. distinct terms / concept	1.37	1.14
<i>Concept distribution</i>		
Single-term concepts (% of concepts)	77.74	88.27
Single-term concepts (% of occ.)	64.44	76.94
Hapax concepts (% of concepts)	37.56	42.68
Hapax concepts (% of occ.)	7.58	11.55
<i>Variation category breakdown (% of occ.)</i>		
NO-VAR	72.28	74.51
LEX	7.97	5.66
RED	9.21	8.93
EXP	1.10	0.39
MORPH	0.59	0.29
GRAPH	7.54	8.73
COMBO	1.32	1.49

Table 4: Terminological variation statistics of the source texts, averaged at the document level.

Table 4 characterizes source-side terminological variation in both corpora. Each concept appears under 1.37 distinct surface forms in ParaNLP and 1.14 in IWSLT2023 on average. 77.7% and 88.3% of concepts respectively are realized by a single form (full distribution in Appendix A.2). Variation is concentrated among frequent concepts, and preferred terms make up 72–75% of occurrences. Among variation

types, RED and GRAPH account for 7–9% each, LEX for 6–8%, and MORPH, EXP, and COMBO together represent less than 3%. Since 38–43% of concepts are hapaxes, any score requiring multiple occurrences is blind to much of the terminology.

The analysis of reference targets in this same setting (Table 5, row *Ref*) confirms that variation is not random noise: most concepts retain their number of distinct surface forms across languages. However, in ParaNLP, 34.4% of concepts exhibit more target-side than source-side realizations, higher than the MT systems. In IWSLT2023 (spoken transcripts), the rate drops to 15.8%. This register asymmetry is consistent with the well-documented tendency of French academic writing to favor lexical variation (*variation élégante*) (Vinay and Darbelnet, 1995).

System	Var.	Ref	E22	E9	Qwen	Llama
ParaNLP						
baseline	=	63.7	69.1	68.5	66.5	66.6
	<	34.4	29.4	29.8	32.2	31.9
base+terms	=	63.8	73.8	72.4	73.2	73.6
	<	34.3	24.1	25.3	24.6	24.8
IWSLT2023						
baseline	=	82.5	67.8	66.7	69.5	64.4
	<	15.8	31.6	32.8	29.9	35.0
base+terms	=	82.5	77.4	74.0	79.7	77.4
	<	15.8	21.5	24.9	19.2	21.5

Table 5: Distribution (%) of concepts by source/target variation (=: EN and FR have the same number of distinct terms; <: FR has more). Row EN>FR omitted (all values below 2.3%). Avg. concepts/doc: 31.48 (ParaNLP), 17.70 (IWSLT2023).

Human translators also routinely replace preferred terms with lexical variants: in the reference, source preferred terms are rendered by a target-side lexical variant (LEX) in 24.98% of cases in IWSLT2023 and 15.68% in ParaNLP (full transfer matrices in Appendix A.3). A concrete case: within a single sentence of the ParaNLP corpus, the English term *reported speech* receives two distinct French translations in the reference (*parole rapportée* and *discours rapporté*, both valid translations of the same phenomenon), a stylistic alternation that a consistency metric may penalize as inconsistency (full ex-

ample with MT outputs in Table 15, Appendix A.4).

MT systems reproduce variation unevenly.

Whether MT outputs reproduce the source’s variation depends on register and variation type. For academic prose (ParaNLP), MT amplification (29–33%) is in the same ballpark as the reference (34.4%); for transcripts (IWSLT2023), MT systems amplify more (30–35%) than the reference (15.8%).

The differences deepen when we break down variation transfer by type, assessing how each source variation type (preferred term, acronym, reduction, etc.) is rendered on the target side (full transfer matrices in Appendix A.3). RED transfers most reliably, with 83–86% preservation on both corpora. GRAPH, in contrast, shows the widest gap between corpora: 50–60% preservation on ParaNLP drops to 14–29% on IWSLT2023, where source acronyms are far more often spelled out as a French phrase than kept as an abbreviation. This likely reflects the translation guidelines of IWSLT2023 rather than the spoken source: even the human reference spells acronyms out more than the MT systems, which a source-side effect could not explain. NO-VAR preservation is likewise higher on ParaNLP (81–85%) than on IWSLT2023 (71–75%). Evaluating all variation types uniformly therefore conflates phenomena with fundamentally different cross-lingual behaviors. For instance, in a single ParaNLP document the acronym *RL* (GRAPH) for *reinforcement learning* is preserved in one context but expanded into different forms in others (full example in Table 16, Appendix A.6).

6.2 Current Metrics Penalize Valid Variation

Method	ParaNLP		IWSLT2023	
	occ	dist	occ	dist
EM	85.0	71.7	90.4	83.3
TCR	86.9	50.5	85.6	53.2
CTV	9.4	12.5	2.7	4.2
EM+TCR	95.1	81.1	96.9	90.0
EM+CTV	91.9	82.6	92.8	87.2
TCR+CTV	88.7	57.7	86.2	55.0
EM+TCR+CTV	96.8	87.9	97.5	91.9

Table 6: Average per-document coverage (%) by method. *occ*: each occurrence; *dist*: each distinct term counted once. TCR uses non-EM variants.

Before reporting scores for the three measurements, we discuss their *coverage*, i.e. the fraction of source-side occurrences each one evaluates. The three evaluation dimensions have complementary coverage profiles (Table 7). EM covers 100% of NO-VAR terms by design but this drops sharply for variants.

EM’s coverage drop on variants is structural: MORPH and COMBO are essentially invisible to glossary-based evaluation, and GRAPH coverage varies across corpora (97% in IWSLT2023, where most acronyms are listed in the glossary, versus 59% in ParaNLP, where many domain-specific acronyms are absent from it). TCR covers 86–87% of occurrences on both corpora but applies only to recurrent canonical forms, assessing each surface form independently and missing cross-variant coherence. CTV has the narrowest raw coverage but provides a qualitatively distinct signal: it is the only dimension that evaluates hapax variants (38–43% of concepts) through their structural relationship to the preferred term. Combined, the three dimensions reach 96.8% of occurrences in ParaNLP and 97.5% in IWSLT2023, leaving a small structurally irreducible residual (Table 6; details in Appendix A.7).

Table 8 reports the three-dimensional scores for MT systems (baseline and base+terms); we focus on the baseline here, leaving base+terms for Section 6.3. Aggregate EM places all systems within a narrow range per corpus, hiding a structural effect: the per-variation breakdown (Appendix A.9) shows that aggregate EM is driven by preferred-term performance (72–85% on NO-VAR) while variant scores are substantially lower. The aggregate therefore says little about how systems handle variants.

Cat.	ParaNLP					IWSLT2023				
	$\bar{N}$	EM	TCR	TCR-EM	CTV	$\bar{N}$	EM	TCR	TCR-EM	CTV
NO-VAR	198	100	90	90	—	88	100	87	87	—
RED	25	43	89	42	32	10	50	85	47	17
MORPH	3	8	2	2	—	1	0	0	0	—
EXP	4	26	32	13	40	4	42	42	42	58
GRAPH	21	59	69	53	32	11	97	82	82	7
LEX	22	35	63	24	32	8	41	46	22	8
COMBO	5	5	26	5	—	4	0	28	0	—

Table 7: % of occurrences evaluated per category, averaged over documents. $\bar{N}$: avg. occurrences. Values rounded to the nearest integer.

Consistency scores expose the cost of penalizing acceptable variation. The reference dominates for

IWSLT2023 under all TCR variants; for ParaNLP, differences narrow. The strictest variant (TCR-EM) results in lower scores across the board, reflecting the compounding difficulty of requiring both accuracy and consistency.

Consistency is hardest to achieve precisely for terms that inherently exhibit variation: TCR scores for LEX and GRAPH are substantially lower than for NO-VAR (Appendix A.9). For instance, the term *pitch* (glossary: *tonie*) receives two contextually appropriate French equivalents (*hauteur, mélodique*) across 14 occurrences in a single ParaNLP document, and a consistency metric severely penalizes this justified variation (full example in Table 17, Appendix A.5).

Metric	ParaNLP					IWSLT2023				
	Ref	E22	E9	Q.	Ll.	Ref	E22	E9	Q.	Ll.
baseline										
EM Acc.	81.5	79.3	79.9	77.8	78.4	72.3	71.3	72.0	74.1	69.9
TCR-EM	83.0	80.8	81.3	79.4	79.8	74.2	73.2	73.5	76.5	72.0
TCR-PR	86.7	88.2	89.2	86.6	89.3	95.1	84.4	87.9	88.1	86.1
TCR-T	85.7	87.2	87.4	85.8	87.7	94.0	83.3	84.0	84.2	82.7
CTV	66.1	72.8	72.7	71.7	74.4	86.7	85.7	85.7	68.2	75.8
base+terms										
EM Acc.	81.5	87.9	85.4	88.3	88.8	72.3	84.5	78.4	87.4	86.6
TCR-EM	83.0	89.0	86.4	89.2	89.7	74.3	86.0	79.6	89.0	87.6
TCR-PR	86.8	91.6	90.5	91.3	90.7	95.1	89.9	90.3	93.0	91.6
TCR-T	85.7	90.7	89.7	90.6	89.8	94.0	87.5	88.3	90.1	89.0
CTV	66.1	67.0	68.5	69.2	68.2	86.7	74.2	81.0	80.6	81.1

Table 8: Scores for baseline and base+terms. E22/E9=Euro22/Euro9, Q.=Qwen, Ll.=Llama, Bold=highest across systems per metric and corpus.

The reference’s CTV score makes the same point from a different angle. While the reference achieves the highest CTV for IWSLT2023, it scores the lowest for ParaNLP. This paradox is explained by the higher target-side variation the reference introduces for ParaNLP, which creates more opportunities for structural coherence failure. Consistency metrics are not stable across definitions. For ParaNLP, each TCR variant ranks a different system as top-1 (Table 8): no single system dominates consistency across all three operationalizations, so the choice of metric is not a neutral implementation detail.

6.3 Glossary Constraints vs. Acceptable Variations

The base+terms block of Table 8 reports MT scores where the preferred terms from the glossary and their expected translations are injected into the translation prompt. Comparing with the baseline, EM and TCR improve substantially for all MT systems: when prompted with the glossary entries, systems readily produce them, showing that they were capable of generating these forms but, without the prompt, often selected alternative wordings. The constraint’s effect on the transfer matrices (Appendix A.3) is selective: NO-VAR → NO-VAR increases by 8–16 points while NO-VAR → LEX drops correspondingly, and other variation types are largely unaffected.

CTV, however, drops for all MT systems under the glossary constraint. Forcing glossary-preferred forms improves both accuracy and consistency at the price of suppressing acceptable target-side variation: the very behavior that Section 6.1 showed to be a hallmark of human translation. This is the central tension exposed by our framework: no single dimension captures the full picture, and a glossary-constrained system that appears to “improve” on EM and TCR may simultaneously regress on CTV. The methodological takeaway is that variation-aware evaluation should condition consistency penalties on whether target-side variation mirrors source-side variation, rather than penalizing all deviation uniformly. Mirroring source variation is a sufficient but not a necessary condition for legitimate target variation: it validates the structural case that CTV measures, but not the stylistic alternation of a term that is itself invariant in the source (e.g., *pitch*, *reported speech*), which has no source counterpart to mirror and can only be adjudicated semantically. CTV is therefore diagnostic rather than prescriptive: it flags where surface consistency and structurally legitimate variation diverge, without licensing every target-side deviation.

7 Conclusion

Terminological variation is a standard property of scientific translation. However, existing automatic metrics systematically penalize it. Analyzing two English–French NLP corpora translated by four medium-size LLMs, we show that human translators introduce more target-side variation than MT systems on academic prose, and that this variation is often semantically motivated. MT systems repro-

duce it unevenly across registers, less the reference for academic prose and more for transcripts.

Current automatic metrics handle this variation poorly: glossary-based EM is dominated by preferred-term performance, translation consistency penalizes the acceptable diversification that human translators produce, and different operationalizations of consistency result in divergent system rankings. Cross-term variation (CTV), the structural coherence dimension we introduce, recovers signal where the other dimensions are blind. No single metric captures terminology quality, motivating variation-aware evaluation that conditions consistency penalties on whether target-side variation mirrors structurally legitimate source-side variation.

The most striking result is that enforcing MT systems to use glossary-preferred forms improves accuracy and consistency but degrades CTV, by suppressing the acceptable target-side variation that the unconstrained baseline preserved. This reveals the need for better MT engines that produce high-quality translations with terminological accuracy, consistency and acceptable variation.

8 Limitations

We carry out our analysis on one language pair (English–French), one domain (NLP scientific text), two corpora, and medium-size LLMs (8B–22B parameters) evaluated under zero-shot conditions. Variation transfer patterns are likely to differ for typologically distant language pairs (e.g., English–Chinese), where acronym and reduction behavior is fundamentally different, and for other specialized domains with distinct terminological profiles. Larger frontier models or translation-specialized systems such as Tower (Alves et al., 2024) may also exhibit different variation behavior; in particular, more capable models might integrate glossary guidance with less suppression of valid variation, an open question that our zero-shot medium-size setting cannot settle. Confirming the generalizability of our findings requires broader experimental coverage.

Our cross-term variation diagnostic (CTV; Dimension 3, Section 4) measures structural coherence under an idealized transfer assumption: it expects each source variation type to be preserved in translation, but legitimate cross-lingual adaptation may well produce a different type (e.g., expanding an acronym

for clarity in the target language).

We do not test how neural or reference-free metrics (e.g., COMET, MetricX, or LLM-as-judge) treat terminological variation. Since they score semantic adequacy rather than surface similarity, they may be more robust to legitimate alternation. However, they commonly operate at the segment level and are not well adapted for the document level. Whether they can also diagnose the terminological inconsistency that our surface-level dimensions target therefore remains an open question.

Several pipeline choices condition our results. The variation classifier uses a closed commercial LLM (gpt-4.1-mini), so exact reproducibility depends on API-version stability; manual validation on a sample of 50 pairs reports 84% type accuracy, with reduced recall on compound variants (11%). As COMBO accounts for fewer than 2% of all occurrences, this affects only finer-grained per-type conclusions. Using an LLM to produce variation labels that are then used to evaluate other LLMs’ translations introduces a potential circular dependency, although the tasks involved (classification vs. translation) are distinct. Alignment errors from Bertalign and SimAlign propagate to all three evaluation dimensions; a manual audit of 100 sampled alignments estimates their rate at 4% (Appendix C). Finally, CONCORDANCER depends on language-specific resources (spaCy, MorphyNet) and rule-based heuristics for dynamic variant detection, which may miss unconventional syntactic constructions. We quantify what CONCORDANCER finds (Appendix A.1) but not what it misses, as no gold-standard annotation of all term occurrences exists to measure detection recall. The framework also requires a bilingual glossary, which may not be available for all domains or language pairs.

Acknowledgements

This work was supported by the French national research agency (ANR) as part of the MaTOS project (grant number: ANR-22-CE23-0033).⁹ 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.

⁹<http://anr-matos.fr/>

References

- Alam, M. M. I., Kvapilíková, I., Anastasopoulos, A., Besacier, L., Dinu, G., Federico, M., Gallé, M., Jung, K., Koehn, P., and Nikoulina, V. (2021). Findings of the WMT shared task on machine translation using terminologies. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussa, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Kocmi, T., Martins, A., Morishita, M., and Monz, C., editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online. Association for Computational Linguistics.
- Alves, D. M., Pombal, J., Guerreiro, N. M., Martins, P. H., Alves, J., Farajian, M. A., Peters, B., Rei, R., Fernandes, P., Agrawal, S., Colombo, P., de Souza, J. G. C., and Martins, A. F. T. (2024). Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.
- Batsuren, K., Bella, G., and Giunchiglia, F. (2021). MorphoNet: a large multilingual database of derivational and inflectional morphology. In Nicolai, G., Gorman, K., and Cotterell, R., editors, *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online. Association for Computational Linguistics.
- Bowker, L. (1998). Variant terminology: Frivolity or necessity? In *Proceedings of the 8th EURALEX International Congress*, pages 487–495, Liège, Belgium.
- Bowker, L. and Hawkins, S. (2006). Variation in the organisation of medical terms. *Terminology*, 12(1):79–110.
- Cárcamo, J. C., Kübler, N., Mestivier, A., and Zhu, L. (2025). Proposition d’une méthodologie pour l’annotation de la variation dénomminative dans les textes scientifiques en sciences de la Terre. In *Beyond Single Words*, Paris, France. Lichao Zhu and Natalie Kübler and Mojca Pecman.
- Čulo, O. and Nitzke, J. (2016). Patterns of terminological variation in post-editing and of cognate use in machine translation in contrast to human translation. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 106–114.
- Daille, B. (2017). *Term variation in specialised corpora: characterisation, automatic discovery and applications.*, volume 19 of *Terminology and Lexicography Research and Practice series*. John Benjamins Publishing Company.
- Deutsch, D., Briakou, E., Caswell, I. R., Finkelstein, M., Galor, R., Juraska, J., Kovacs, G., Lui, A., Rei, R., Riesa, J., Rijhwani, S., Riley, P., Salesky, E., Trabelsi, F., Winkler, S., Zhang, B., and Freitag, M. (2025). WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.
- Dinu, G., Mathur, P., Federico, M., and Al-Onaizan, Y. (2019). Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.
- Fernández-Silva, S. and Kerremans, K. (2011). Terminological variation in source texts and translations: A pilot study. *Meta*, 56(2):318–335.
- Gašpar, A., Seljan, S., and Kučič, V. (2022). Measuring Terminology Consistency in Translated Corpora: Implementation of the Herfindahl-Hirshman Index. *Information*, 13(2).
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., and et al., B. C. (2024). The Llama 3 Herd of Models. Preprint arXiv:2407.21783.
- Hokamp, C. and Liu, Q. (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.
- Honnibal, M., Montani, I., van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. <https://spacy.io>.
- Hu, J. E., Rudinger, R., Post, M., and Van Durme, B. (2019). PARABANK: Monolingual bitext generation

- and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6521–6528.
- ibn Alam, M. M., Anastasopoulos, A., Besacier, L., Cross, J., Gallé, M., Koehn, P., and Nikoulina, V. (2021). On the evaluation of machine translation for terminology consistency. Preprint arXiv:2106.11891.
- ISO (2019). ISO 1087:2019: Terminology work and terminology science. Vocabulary. International Organization for Standardization, Clause 3.4.
- Itagaki, M., Aikawa, T., and He, X. (2007). Automatic validation of terminology translation consistency with statistical method. In Maegaard, B., editor, *Proceedings of Machine Translation Summit XI: Papers*, Copenhagen, Denmark.
- Jalili Sabet, M., Dufter, P., Yvon, F., and Schütze, H. (2020). SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.
- Juraska, J., Deutsch, D., Finkelstein, M., and Freitag, M. (2024). MetricX-24: The Google submission to the WMT 2024 metrics shared task. In Haddow, B., Kocmi, T., Koehn, P., and Monz, C., editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.
- Kocmi, T. and Federmann, C. (2023). GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.
- Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. (2023). Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Liu, L. and Zhu, M. (2022). Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.
- Lu, Q., Ding, L., Zhang, K., Zhang, J., and Tao, D. (2025). MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587, Abu Dhabi, UAE. Association for Computational Linguistics.
- Lyu, X., Li, J., Gong, Z., and Zhang, M. (2021). Encouraging lexical translation consistency for document-level neural machine translation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3265–3277, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Martins, P. H., Alves, J., Fernandes, P., Guerreiro, N. M., Rei, R., Farajian, A., Klimaszewski, M., Alves, D. M., Pombal, J., Boizard, N., Faysse, M., Colombo, P., Yvon, F., Haddow, B., de Souza, J. G. C., Birch, A., and Martins, A. F. T. (2025). EuroLLM-9B: Technical Report. Preprint arXiv:2506.04079.
- Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.
- Pecman, M. (2014). Variation as a cognitive device: How scientists construct knowledge through term formation. *Terminology*, 20(1):1–24.
- Peng, Z., Zhu, L., Bawden, R., Bénard, M., de la Clergerie, É., Huguin, M., Kübler, N., Lerner, P., Mestivier, A., and Yvon, F. (2026). Parallel Corpora of Scholarly Documents for English-French Machine Translation. In *Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC)*, Palma de Mallorca, Spain. LREC2026.
- Post, M. and Vilar, D. (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, New Orleans, Louisiana. Association for Computational Linguistics.

- Ramos, M. M., Alves, D. M., Gisserot-Boukhlef, H., Almes, J., Martins, P. H., Fernandes, P., Pombal, J., Guerreiro, N. M., Rei, R., Boizard, N., Farajian, A., Klimaszewski, M., de Souza, J. G. C., Haddow, B., Yvon, F., Colombo, P., Birch, A., and Martins, A. F. T. (2026). Eurollm-22b: Technical report. Preprint arXiv:2602.05879.
- Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., C. de Souza, J. G., Glushkova, T., Alves, D., Coheur, L., Lavie, A., and Martins, A. F. T. (2022). CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Salesky, E., Darwish, K., Al-Badrashiny, M., Diab, M., and Niehues, J. (2023). Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In Salesky, E., Federico, M., and Carpuat, M., editors, *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online). Association for Computational Linguistics.
- Semenov, K. and Bojar, O. (2022). Automated evaluation metric for terminology consistency in MT. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Névéol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation*
- (WMT), pages 450–457, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Semenov, K., Huang, X., Zouhar, V., Berger, N., Zhu, D., Oncevay, A., and Chen, P. (2025). Findings of the WMT25 terminology translation task: Terminology is useful especially for good MTs. In Haddow, B., Kocmi, T., Koehn, P., and Monz, C., editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 554–576, Suzhou, China. Association for Computational Linguistics.
- Semenov, K., Zouhar, V., Kocmi, T., Zhang, D., Zhou, W., and Jiang, Y. E. (2023). Findings of the WMT 2023 shared task on machine translation with terminologies. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.
- Toral, A. (2019). Post-editsese: an exacerbated translationese. In Forcada, M., Way, A., Haddow, B., and Senrich, R., editors, *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.
- Vanmassenhove, E., Shterionov, D., and Gwilliam, M. (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. Association for Computational Linguistics.
- Vanmassenhove, E., Shterionov, D., and Way, A. (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.
- Vinay, J.-P. and Darbelnet, J. (1995). *Comparative Stylistics of French and English: A Methodology for Translation*. John Benjamins, Amsterdam. Translated and edited by Juan C. Sager and M.-J. Hamel. Originally published as *Stylistique comparée du français et de l’anglais* (1958).
- Wang, Y., Zeng, J., Liu, X., Wong, D. F., Meng, F., Zhou, J., and Zhang, M. (2025). DelTA: An online document-level translation agent based on multi-level memory. In

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., and et al., J. Y. (2025). Qwen3 technical report. Preprint arXiv:2505.09388.

A Appendix

A.1 Term Detection Coverage

	ParaNLP	IWSLT2023
Exact-form match (%)	61.9%	62.7%
Lemma match (%)	82.2%	86.6%
Avg. terms / doc	273.8	120.3

Table 9: CONCORDANCER recognition coverage, averaged per document. Exact-form and lemma matching are baselines that a naive identification method could use; CONCORDANCER recovers the remaining terms by detecting further variants.

	ParaNLP		IWSLT2023	
	ING	OOG	ING	OOG
Avg. terms / doc	232.8	40.9	109.1	11.2
NO-VAR	84.3%	–	82.5%	–
GRAPH	7.0%	11.3%	9.2%	3.9%
LEX	3.5%	35.7%	3.2%	31.3%
EXP	0.5%	5.0%	0.3%	1.7%
RED	4.6%	36.3%	4.8%	45.5%
MORPH	0.0%	4.1%	0.0%	2.7%
COMBO	0.1%	7.6%	0.0%	14.9%

Table 10: Distribution of variation types among CONCORDANCER-detected terms, averaged per document and split by whether the term’s lemma is listed in the glossary (*ING*, in-glossary) or not (*OOG*, out-of-glossary). Percentages are within each group.

Table 9 reports the recognition coverage of CONCORDANCER on both corpora. Exact-form and lemma matching (the two naive identification baselines) would respectively retrieve 61.9% and 82.2% of terms for ParaNLP, and 62.7% and 86.6% for

IWSLT2023. CONCORDANCER, which detects additional variant types, identifies the remaining terms.

Table 10 breaks down CONCORDANCER-detected terms into *in-glossary* terms (*ING*), whose lemma appears in the glossary, and *out-of-glossary* terms (*OOG*), discovered by CONCORDANCER beyond its scope. For *ING* terms, more than 80% are preferred terms (NO-VAR). For *OOG* terms, the distribution shifts: LEX and RED together account for more than 70%, meaning CONCORDANCER primarily discovers out-of-glossary terms through lexical and reduction variants. Since no glossary reference exists for *OOG* terms, their translation cannot be assessed by term-level accuracy (Dimension 1), motivating the complementary evaluation dimensions.

A.2 Distribution of Distinct Terms per Concept

Figure 1 reports the distribution of the number of distinct canonical surface forms (preferred term and variants) per concept, averaged at the document level. The distribution is strongly skewed: most concepts are realized by a single form in both corpora, with a long tail of concepts showing two or more variants.

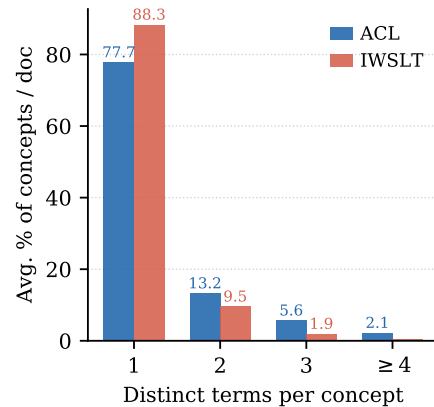


Figure 1: Distribution of the number of distinct terms per concept, averaged by document, showing the proportion of concepts with exactly k distinct terms.

A.3 Variation Transfer Matrices

Tables 11 and 12 present the full source-to-target variation transfer matrices under the baseline condition. Each cell reports the percentage of source occurrences of a given variation type (rows) that are translated as a given target variation type (columns).

Diagonal values indicate type preservation. Key observations beyond those discussed in Section 6.1:

- **NO-VAR** → **NO-VAR**: 81–85% on ParaNLP but only 71–75% on IWSLT2023. The off-diagonal NO-VAR → LEX is the main source of preferred-term variation: 10–12% on ParaNLP and 18–25% on IWSLT2023, with the reference showing the highest rate on both corpora.
- **GRAPH** → **GRAPH**: 50–60% on ParaNLP but only 14–29% on IWSLT2023, where 63–79% of source GRAPH are realized as RED instead. This likely reflects the translation guidelines of IWSLT2023, where source acronyms are systematically spelled out rather than kept as an abbreviation (the human reference does so even more than the MT systems), rather than the spoken origin of the source.
- **RED** → **RED**: 83–86% on both corpora, the most stable transfer across all types and systems.
- **LEX** → **LEX**: 58–63% on ParaNLP and 52–77% on IWSLT2023, with the reference achieving notably higher preservation than MT systems on IWSLT2023.

Tables 13 and 14 present the same matrices under the base+terms condition. The glossary constraint selectively affects preferred-term transfer: NO-VAR → NO-VAR increases from 81–84% to 90–93% on ParaNLP and from 71–75% to 79–91% on IWSLT2023, while NO-VAR → LEX drops from 10–12% to 3–5% on ParaNLP and from 18–23% to 5–16% on IWSLT2023. Other variation types are largely unaffected: RED → RED remains at 84–86% and GRAPH → GRAPH at 49–54% on ParaNLP, indicating that the glossary constraint acts primarily on preferred terms and does not propagate to variant categories.

A.4 Legitimate Variation Example: *reported speech*

Table 15 shows translations of the sentence below by all systems. Within a single sentence, the reference alternates between *parole rapportée* and *discours rapporté* as renderings of *reported speech*, a stylistic choice that a consistency metric would penalize as inconsistency, although both forms are standard French

equivalents of the same linguistic phenomenon. All four baseline MT systems instead produce a single form repeated.

A.5 Legitimate Variation Example: *pitch*

Table 17 shows three of the 14 occurrences of *pitch* (glossary: *tonie*) in a single ParaNLP document, translated as either *hauteur* (for pitch values) or *mélodique* (in compound expressions such as *contour mélodique* or *patron mélodique*). All renderings are contextually appropriate, yet a consistency metric severely penalizes the alternation.

A.6 Variation Type Variability Example: *reinforcement learning*

Table 16 shows three occurrences of the acronym *RL* (a GRAPH variant of *reinforcement learning*) in a single ParaNLP document. The acronym is preserved in one context but expanded into different forms in others, illustrating the within-document variability of GRAPH transfer.

Source	Reference
... pitch values were ex- tracted and averaged...	...les valeurs de hauteur ont été extraites...
...more pitch movement on penultimate syllables...	...plus de mouvement mélodique sur la pénul- tième...
...a specific pitch pattern on the penultimate syllable.	...un schéma mélodique spécifique sur la syllabe pénultième.

Table 17: The term *pitch* (glossary: *tonie*) translated as 2 distinct lexical variants (*hauteur*, *mélodique*) across 14 occurrences by the Reference. All translations are contextually appropriate, yet a consistency metric results in a low score for this term.

A.7 Unevaluated Terms

Tables 18 and 19 characterize the terms that remain unevaluated by any of the three dimensions. For ParaNLP, 3.2% of occurrences (12.1% of distinct terms) are never evaluated; for IWSLT2023, 2.5% (8.1%). These residual terms are hapax occurrences absent from the glossary and without a CTV-eligible relationship to a preferred term, representing a small but structurally irreducible gap.

System	Src	NO-VAR	LEX	RED	GRAPH	EXP	MORPH	COMBO	∅
Ref.	NO-VAR	85.2	9.9	1.5	0.9	0.1	2.4	0.0	0.1
	LEX	7.1	63.0	18.7	4.1	1.8	4.1	0.4	0.9
	RED	0.1	7.4	86.2	2.4	0.6	1.4	0.2	1.7
	GRAPH	13.2	13.6	9.6	49.9	5.4	3.9	3.1	1.1
	EXP	9.0	14.4	9.6	13.9	26.1	16.4	7.9	2.8
	MORPH	12.3	24.1	17.6	0	10.2	35.8	0	0
	COMBO	7.2	34.2	14.0	13.4	5.7	22.7	1.9	0.9
Euro22	NO-VAR	83.0	10.9	1.6	0.7	0.1	3.6	0	0.0
	LEX	6.8	59.5	22.9	4.8	0.9	3.9	0	1.1
	RED	0	9.7	83.4	2.2	0.2	2.7	0.1	1.6
	GRAPH	9.5	9.9	8.4	53.9	10.4	2.4	4.8	0.7
	EXP	4.6	19.3	1.5	24.8	31.4	7.2	8.4	2.8
	MORPH	13.3	14.0	17.6	0	8.9	44.5	1.1	0.6
	COMBO	4.4	26.4	14.1	11.7	8.5	31.0	2.9	0.9
Euro9	NO-VAR	83.5	10.4	1.6	0.9	0.1	3.4	0	0.1
	LEX	6.2	61.8	21.5	4.8	1.2	3.2	0	1.3
	RED	0.2	10.2	82.8	2.5	0.3	2.4	0	1.6
	GRAPH	6.6	7.9	7.6	58.9	7.4	4.3	6.3	0.9
	EXP	8.5	15.5	9.9	23.4	24.6	13.9	1.4	2.8
	MORPH	13.8	19.9	15.3	0	7.2	42.6	0.6	0.6
	COMBO	4.0	25.7	16.3	8.7	7.3	34.7	2.5	0.9
Qwen	NO-VAR	81.4	12.5	1.4	1.7	0.1	2.8	0.0	0.1
	LEX	6.3	58.8	24.9	4.0	1.6	3.1	0.1	1.3
	RED	0.1	10.2	83.0	2.3	0.4	2.2	0.1	1.8
	GRAPH	5.9	13.7	7.5	57.5	7.1	0.5	7.2	0.7
	EXP	2.1	23.8	2.3	24.2	18.8	17.9	8.2	2.8
	MORPH	14.6	20.5	16.7	0.6	5.7	40.9	0.6	0.6
	COMBO	5.9	30.8	11.6	7.3	6.8	34.4	2.2	0.9
Llama	NO-VAR	81.6	11.1	1.2	2.5	0.1	3.4	0	0.0
	LEX	6.0	59.6	22.2	5.0	1.7	4.0	0.2	1.3
	RED	0.1	8.7	85.5	2.3	0.2	1.5	0	1.7
	GRAPH	7.7	7.4	10.5	59.4	5.0	1.9	7.3	0.9
	EXP	9.2	18.9	2.7	23.4	32.9	7.9	2.1	2.8
	MORPH	10.0	27.3	12.7	0	11.4	38.1	0	0.6
	COMBO	4.0	33.4	15.1	8.1	7.3	29.1	2.2	0.9

Table 11: Variation transfer from source to target (% , mean per document) per system, baseline condition, ParaNLP. Rows: source variation category per system; columns: target variation category. Diagonal (preservation) in bold. Source distribution (ref., mean/doc): NO-VAR: 72.3% (198), LEX: 7.9% (22), RED: 9.0% (25), GRAPH: 7.7% (21), EXP: 1.1% (4), MORPH: 0.6% (2), COMBO: 1.3% (5).

System	Src	NO-VAR	LEX	RED	GRAPH	EXP	MORPH	COMBO	∅
Ref	NO-VAR	72.6	21.1	1.4	0.8	0	4.0	0	0.1
	LEX	3.5	77.2	8.1	0.6	0	10.6	0	0
	RED	0	11.5	86.4	0	0	1.0	1.1	0
	GRAPH	0	0	78.6	14.2	2.0	5.3	0	0
	EXP	0	41.7	0	0	58.3	0	0	0
	MORPH	0	0	66.7	0	0	33.3	0	0
	COMBO	0	8.3	20.8	8.3	8.3	45.8	8.3	0
Euro22	NO-VAR	72.2	19.8	1.0	0.6	0	6.2	0	0.1
	LEX	11.5	54.7	13.7	9.5	0	10.6	0	0
	RED	0	15.7	81.2	0	0	2.0	1.1	0
	GRAPH	0	0	66.7	26.6	1.0	5.3	0.4	0
	EXP	0	25.0	0	16.7	58.3	0	0	0
	MORPH	0	0	100.0	0	0	0	0	0
	COMBO	0	18.1	37.5	8.3	8.3	19.4	8.3	0
Euro9	NO-VAR	73.1	19.2	1.2	1.0	0	5.6	0	0
	LEX	14.9	58.6	8.1	7.3	0	11.1	0	0
	RED	0	10.4	86.5	0	1.1	2.0	0	0
	GRAPH	0.4	0	66.7	26.6	1.0	5.3	0	0
	EXP	0	0	41.7	0	0	8.3	50.0	0
	MORPH	0	0	66.7	33.3	0	0	0	0
	COMBO	0	2.8	36.1	16.7	8.3	27.8	8.3	0
Qwen	NO-VAR	74.8	17.3	0.6	0.9	0	4.7	0.2	1.3
	LEX	14.9	52.2	12.6	9.8	0	10.6	0	0
	RED	0	13.4	84.8	0	0	0	0	1.7
	GRAPH	1.7	1.8	65.0	25.2	1.0	5.0	0	0.3
	EXP	0	41.7	0	0	0	0	58.3	0
	MORPH	0	0	66.7	33.3	0	0	0	0
	COMBO	0	18.1	30.6	8.3	8.3	18.1	16.7	0
Llama	NO-VAR	70.8	20.2	1.4	1.4	0	6.0	0.1	0.1
	LEX	11.0	56.7	11.4	9.7	0	11.2	0	0
	RED	0	12.9	85.1	0	0	2.0	0	0
	GRAPH	0.4	1.4	62.9	29.0	1.0	4.3	1.0	0
	EXP	0	75.0	0	16.7	0	8.3	0	0
	MORPH	0	0	66.7	0	0	0	33.3	0
	COMBO	0	18.1	30.6	8.3	8.3	18.1	16.7	0

Table 12: Variation transfer from source to target (% , mean per document) per system, baseline condition, IWSLT2023. Rows: source variation category per system; columns: target variation category. Diagonal (preservation) in bold. Source distribution (ref., mean/doc): NO-VAR: 74.5% (88), LEX: 5.7% (7), RED: 8.9% (10), GRAPH: 8.7% (11), EXP: 0.4% (4), MORPH: 0.3% (1), COMBO: 1.5% (4).

System	Src	NO-VAR	LEX	RED	GRAPH	EXP	MORPH	COMBO	∅
Ref.	NO-VAR	85.2	9.9	1.5	0.9	0.1	2.4	0.0	0.1
	LEX	7.1	63.0	18.7	4.1	1.8	4.1	0.4	0.9
	RED	0.1	7.4	86.2	2.4	0.6	1.4	0.2	1.7
	GRAPH	13.2	13.6	9.6	49.9	5.4	3.9	3.1	1.1
	EXP	9.0	14.4	9.6	13.9	26.1	16.4	7.9	2.8
	MORPH	12.3	24.1	17.6	0	10.2	35.8	0	0
	COMBO	7.2	34.2	14.0	13.4	5.7	22.7	1.9	0.9
Euro22	NO-VAR	93.0	3.3	1.6	0.5	0.1	1.3	0.0	0.0
	LEX	8.4	57.4	23.3	4.8	0.7	4.2	0	1.1
	RED	0.2	8.7	85.0	2.3	0.2	2.0	0.1	1.6
	GRAPH	15.5	6.7	11.0	50.0	9.2	4.1	2.6	0.9
	EXP	9.7	16.2	2.7	23.2	34.0	9.3	2.1	2.8
	MORPH	18.9	8.3	15.3	0	7.8	47.9	1.1	0.6
	COMBO	9.1	30.3	9.7	8.7	6.4	31.2	3.8	0.9
Euro9	NO-VAR	90.2	5.3	1.4	0.7	0.1	2.2	0	0.1
	LEX	7.7	60.0	21.8	3.6	2.0	3.1	0.1	1.7
	RED	0.1	9.0	84.1	2.4	0.4	2.3	0.1	1.6
	GRAPH	14.8	5.7	9.9	50.9	9.8	3.2	4.4	1.3
	EXP	11.3	13.3	11.7	26.6	27.1	6.9	0	3.0
	MORPH	19.5	14.6	13.8	0.6	7.2	41.7	0.6	2.1
	COMBO	7.9	25.8	19.5	7.8	7.3	28.3	2.5	1.0
Qwen	NO-VAR	93.5	3.2	1.5	0.5	0.1	1.1	0	0.1
	LEX	10.5	53.5	25.2	4.6	1.5	3.3	0	1.3
	RED	0.7	7.3	85.1	2.2	0.3	2.7	0	1.7
	GRAPH	18.2	8.5	9.8	49.1	9.3	2.6	1.8	0.7
	EXP	6.0	16.5	2.1	17.0	44.4	10.7	0.6	2.8
	MORPH	24.1	10.8	14.2	0	12.9	36.9	0.6	0.6
	COMBO	12.9	26.2	11.7	12.8	7.2	25.9	2.5	0.9
Llama	NO-VAR	93.2	3.4	1.4	0.7	0.1	1.1	0	0.1
	LEX	8.8	54.7	23.4	5.8	1.7	4.2	0	1.4
	RED	0.6	7.3	86.0	2.4	0.3	1.7	0	1.7
	GRAPH	15.4	6.1	12.2	54.4	6.4	1.4	3.4	0.7
	EXP	12.2	10.9	2.7	31.9	36.5	3.0	0	2.8
	MORPH	25.9	19.7	10.4	0	9.5	33.9	0	0.6
	COMBO	9.1	32.1	10.8	8.7	8.2	26.9	3.3	0.9

Table 13: Variation transfer from source to target (% , mean per document) per system, base+terms condition, ParaNLP. Rows: source variation category per system; columns: target variation category. Diagonal (preservation) in bold. Source distribution (ref., mean/doc): NO-VAR: 72.3% (198), LEX: 7.9% (22), RED: 9.0% (25), GRAPH: 7.7% (21), EXP: 1.1% (4), MORPH: 0.6% (2), COMBO: 1.3% (5).

System	Src	NO-VAR	LEX	RED	GRAPH	EXP	MORPH	COMBO	$\emptyset$
Ref.	NO-VAR	72.6	21.1	1.4	0.8	0	4.0	0	0.1
	LEX	3.5	77.2	8.1	0.6	0	10.6	0	0
	RED	0	11.5	86.4	0	0	1.0	1.1	0
	GRAPH	0	0	78.6	14.2	2.0	5.3	0	0
	EXP	0	41.7	0	0	58.3	0	0	0
	MORPH	0	0	66.7	0	0	33.3	0	0
	COMBO	0	8.3	20.8	8.3	8.3	45.8	8.3	0
Euro22	NO-VAR	86.7	7.5	1.4	0.7	0.1	3.5	0.1	0
	LEX	21.6	40.3	12.6	8.5	0	16.9	0	0
	RED	1.2	10.4	84.1	0	2.2	2.0	0	0
	GRAPH	12.1	0	65.3	21.2	1.4	0	0	0
	EXP	0	91.7	0	0	8.3	0	0	0
	MORPH	0	0	100.0	0	0	0	0	0
	COMBO	8.3	19.4	36.1	8.3	8.3	11.1	8.3	0
Euro9	NO-VAR	80.4	12.8	1.1	0.4	0	5.0	0.1	0.1
	LEX	16.6	52.3	11.4	7.8	0	11.8	0	0
	RED	0	10.3	86.6	0	1.1	2.0	0	0
	GRAPH	8.5	0	63.9	26.6	1.0	0	0	0
	EXP	0	0	41.7	8.3	0	0	50.0	0
	MORPH	0	33.3	33.3	33.3	0	0	0	0
	COMBO	8.3	20.8	36.1	8.3	8.3	9.7	8.3	0
Qwen	NO-VAR	89.8	5.2	1.1	0.7	0	2.7	0.1	0.4
	LEX	22.1	43.1	9.3	6.0	0	13.1	0	6.4
	RED	3.0	9.2	85.6	0	1.1	1.1	0	0
	GRAPH	8.4	0	65.4	23.2	1.0	2.0	0	0
	EXP	8.3	25.0	0	8.3	58.3	0	0	0
	MORPH	0	0	33.3	33.3	0	0	0	33.3
	COMBO	8.3	2.8	36.1	8.3	8.3	11.1	25.0	0
Llama	NO-VAR	88.9	6.6	1.5	0.7	0.1	2.2	0	0
	LEX	26.2	40.7	10.5	4.8	0	17.8	0	0
	RED	1.2	9.8	86.6	0	1.1	1.2	0	0
	GRAPH	13.7	0	58.7	26.6	1.0	0	0	0
	EXP	0	25.0	0	16.7	58.3	0	0	0
	MORPH	0	0	66.7	0	0	0	33.3	0
	COMBO	8.3	0	38.9	0	8.3	36.1	8.3	0

Table 14: Variation transfer from source to target (% , mean per document) per system, base+terms condition, IWSLT2023. Rows: source variation category per system; columns: target variation category. Diagonal (preservation) in bold. Source distribution (ref., mean/doc): NO-VAR: 74.5% (88), LEX: 5.7% (7), RED: 8.9% (10), GRAPH: 8.7% (11), EXP: 0.4% (4), MORPH: 0.3% (1), COMBO: 1.5% (4).

A.8 TCR Metric Divergence Example

Table 20 illustrates how the three TCR metrics diverge on the same data. For 5 occurrences of *neural network* (glossary: *réseau de neurones*), TCR-EM scores 0.40, TCR-PR 0.60, and TCR-T 0.50.

	Translation	EM	PR	Temporal
o_1	réseau neuronal	×	=	
o_2	réseau neuronal	×	=	=
o_3	réseau de neurones	✓	≠	≠
o_4	réseau de neurones	✓	≠	=
o_5	réseau neuronal	×	=	≠
Score		2/5	3/5	2/4

Table 20: Translations of *neural network* (glossary: *réseau de neurones*). **EM**: glossary match; **PR**: agrees with pseudo-ref (o_1); **Temporal**: agrees with previous occ.

A.9 Scores Breakdown by Variation Label

A.9.1 Baseline Condition

Term-level accuracy (EM). Table 21 reports EM Micro broken down by source variation category. EM ranges from 70.0–74.1 on IWSLT2023 and 77.8–81.5 on ParaNLP, with Qwen leading on IWSLT2023 (74.05) and Ref on ParaNLP (81.52). NO-VAR terms achieve 72–85% EM accuracy across systems. Performance drops sharply for variants: LEX reaches only 34–55%, GRAPH 42–74%, and MORPH and COMBO score 0% across all systems on both corpora. Since NO-VAR accounts for 72–74% of occurrences, the aggregate EM is structurally dominated by preferred-term performance.

Translation consistency (TCR). Tables 22, 23, and 24 report the three TCR variants broken down by variation category. Under the least restrictive variant (TCR-PR), the reference dominates on IWSLT2023 (95.09 vs. 84.4–88.1 for MT systems) but differences narrow on ParaNLP (86.8–89.3). The per-variation pattern is consistent across metrics: NO-VAR terms score substantially higher (88–95%) than LEX (47–55% on IWSLT2023, 78–88% on ParaNLP) or GRAPH (53–89% depending on corpus and system). COMBO scores are low but non-zero (6–24%), indicating that the few compound variants with $k \geq 2$ occurrences are rarely translated consistently.

A notable observation concerns system ranking

stability. On ParaNLP under baseline, each TCR variant ranks a different system as top-1: TCR-EM favors the reference (82.99), TCR-PR favors Llama (89.31), and TCR-T favors E9 (87.93). No single system dominates consistency across all operationalizations, indicating that the choice of consistency metric is not neutral and can substantially affect system ranking.

Cross-term variation (CTV). Table 25 reports variation preservation by source variation category. On IWSLT2023, the reference achieves the highest CTV (86.67), outperforming all MT systems (68–86%). On ParaNLP, the pattern reverses: the reference scores lowest (66.06) while Llama leads (74.63). This asymmetry is explained by the target-side amplification documented in Section 6.1.

Per variation type, RED preservation is consistently high across all systems (82–100% on IWSLT2023, 81–87% on ParaNLP), confirming that reductions transfer reliably. GRAPH coherence varies more: the reference achieves 90% on IWSLT2023 but only 58% on ParaNLP, while MT systems range from 50–86% on IWSLT2023 and 42–78% on ParaNLP. LEX coherence shows the widest system-dependent spread (50–86% on IWSLT2023, 57–72% on ParaNLP), and EXP is both rare and variable.

A.9.2 Glossary-Constrained Condition (base+terms)

Term-level accuracy (EM). Under base+terms (Table 26), EM Micro increases substantially for all MT systems: +10–17 points on IWSLT2023 (from 70–74% to 78–90%) and +6–10 points on ParaNLP (from 78–80% to 85–89%), confirming that baseline systems already possess the correct terms but do not always produce them without guidance. The improvement is concentrated on NO-VAR terms (+14–18 points on IWSLT2023, +8–12 points on ParaNLP), with smaller gains on RED and GRAPH. MORPH and COMBO remain at 0%.

Translation consistency (TCR). TCR metrics also improve under base+terms (Tables 27, 28, 29). On NO-VAR terms, TCR-PR rises from 86–88% to 91–95% on IWSLT2023 and from 88–90% to 93–95% on ParaNLP, indicating that forcing preferred-term usage reduces translation inconsistency. The improvement is smaller for variant categories, which are not directly targeted by the glossary constraint.

Cross-term variation (CTV). Under base+terms (Table 30), CTV drops for most MT systems compared to baseline, particularly on ParaNLP (from 72–75% to 67–69%). On IWSLT2023, GRAPH coherence improves slightly for some systems (Qwen: 50% → 90%), but RED coherence decreases (95–100% → 83–96%). This confirms the tension identified in the main analysis: the glossary constraint improves accuracy and consistency at the cost of suppressing the target-side variation that CTV measures. The reference, unchanged across conditions, serves as a stable baseline.

A.10 Top-5 Most Frequent Concepts

Table 31 provides a detailed breakdown of the five most frequent concepts per document. The five most frequent concepts account for a cumulative 47.2% of all occurrences in ParaNLP and 52.5% in IWSLT2023, with up to 2.38 distinct terms per concept. Notably, the proportion of preferred-term usage (%NO-VAR) decreases from Top-1 to Top-5 (e.g., 87.3% to 60.3% on ParaNLP), suggesting that the most frequent concept in a document tends to be expressed more canonically, while less central concepts allow for greater surface variation.

Rank	ParaNLP			IWSLT2023		
	%occ	Terms	%NO-VAR	%occ	Terms	%NO-VAR
Top-1	16.8%	1.66	87.3%	19.4%	1.30	88.7%
Top-2	10.7%	2.19	70.2%	10.7%	1.20	92.1%
Top-3	8.2%	2.03	77.7%	8.7%	1.40	65.0%
Top-4	6.4%	1.88	73.2%	7.5%	1.30	73.9%
Top-5	5.1%	2.38	60.3%	6.2%	1.30	69.0%

Table 31: Average statistics for the top-5 most frequent concepts per document. *%occ* = share of total occurrences in the document; *Terms* = average number of distinct surface forms; *%NO-VAR* = percent-age of occurrences using the preferred term.

A.11 Model Cards

We list below the HuggingFace model cards for the four open-weight LLMs used in the MT experiments.

- Llama3.1-8B-Instruct: <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>
- Qwen3-8B: <https://huggingface.co/Qwen/Qwen3-8B>

¹⁰<https://huggingface.co/Unbabel/wmt22-cometkiwi-da>

¹¹<https://huggingface.co/google/metricx-24-hybrid-large-v2p6>

- EuroLLM-9B-Instruct: <https://huggingface.co/utter-project/EuroLLM-9B-Instruct>

- EuroLLM-22B-Instruct: <https://huggingface.co/utter-project/EuroLLM-22B-Instruct-2512>

A.12 ParaNLP Corpus Construction Details

The comparable articles in the ParaNLP corpus were transformed into parallel data through automatic sentence segmentation and alignment. Poorly aligned or unaligned sentences were handled via automatic post-editing (APE) (4,928 sentences) and machine translation (1,381 sentences), resulting in 6,272 parallel sentences in total. Three articles, including 199 MT sentences and 42 APE sentences, were further post-edited by a native French speaker with expertise in NLP.

A.13 Global Score Evaluation

For completeness, we report general MT quality scores (COMET-Kiwi (Rei et al., 2022)¹⁰ and MetricX-QE (Juraska et al., 2024)¹¹) for each system on both corpora. These complement the terminology-focused analysis in the main paper by situating each system’s overall translation quality.

System	COMET-Kiwi ↑	MetricX-QE ↓
EuroLLM-22B-greedy	0.8380	2.5987
EuroLLM-9B-greedy	0.8390	2.5797
Llama-3.1-8B-greedy	0.8286	2.8116
Qwen3-8B-top-p	0.8344	2.7076
Reference	0.8242	2.9656

Table 32: Global evaluation scores for the ParaNLP corpus, where each global score is computed as the mean of document-level scores, and each document-level score is the mean of its sentence-level scores, using COMET-Kiwi (↑) and MetricX-QE (↓).

System	COMET-Kiwi $\uparrow$	MetricX-QE $\downarrow$
EuroLLM-22B-greedy	0.8229	2.9226
EuroLLM-9B-greedy	0.8229	2.8744
Llama-3.1-8B-greedy	0.8067	3.4546
Qwen3-8B-top-p	0.8151	3.3200
Reference	0.8030	3.3728

Table 33: Global evaluation scores for the IWSLT2023 coprus, where each global score is computed as the mean of document-level scores, and each document-level score is the mean of its sentence-level scores, using COMET-Kiwi ($\uparrow$) and MetricX-QE ($\downarrow$).

Results Analysis Gobal Score. Across both corpora, all MT systems consistently outperform the human reference under both a known artefact of QE metrics, which are predominantly trained on MT outputs and thus tend to assign systematically lower scores to human-generated translations (Deutsch et al., 2025). Importantly, scores were computed segment-by-segment, averaged to document level, then aggregated per system, meaning *no contextual information was available during evaluation*. So terminological consistency is therefore not captured, which likely flatters MT systems over human translators who actively manage such consistency across documents. System rankings remain stable throughout: **EuroLLM-9B-greedy** achieves the best or tied-best MetricX-QE in every setting, notably surpassing its 22B counterpart, while **Llama-3.1-8B-greedy** consistently ranks last. Scores are uniformly higher on ParaNLP than on IWSLT2023, reflecting the greater regularity of academic prose versus spontaneous speech.

A.14 Variation Classification Typology

This appendix provides a comprehensive reference for the terminological variation typology used throughout this work. The typology distinguishes between single-type variations (GRAPH, MORPH, RED, EXP, LEX) and multiple-type variations (COMBO), which combine two or more single types. Table 34 provides a complete reference of all variation types and subtypes.

B Manual Evaluation of LLM-based Variation Labeling

The variation labels assigned to lexicalised alternative terms (Section 3) are produced by a ChatGPT-

based classifier with few-shot prompting (see Appendix 9). To assess the reliability of this automatic labeling, we conducted a manual evaluation on a sample of 50 randomly selected (preferred term, variant) pairs drawn from the full set of labeled instances. The evaluation was restricted to pairs where the preferred term and the variant differ, that is, pairs not automatically labeled by CONCORDANCER on the basis of structural patterns. This ensures that the evaluation targets exclusively the LLM classifier’s judgment, independently of the rule-based component of the pipeline. For each sampled pair, the predicted variation type and subtype (e.g., GRAPH acronym, MORPH constituent reordering) were compared against gold labels assigned by one of the authors, a native French speaker fluent in English, with expertise in computational linguistics. Tables 35 and 36 report per-class precision and recall at the type and subtype levels respectively. The 49 evaluated pairs are listed in Table 37 (one pair excluded as the variant was not a genuine terminological link).

Type	Support	Precision	Recall	Accuracy
GRAPH	14	67%	100%	—
MORPH	1	100%	100%	—
RED	4	100%	100%	—
EXP	2	100%	100%	—
LEX	19	95%	100%	—
COMBO	9	100%	11%	—
Total	49	94%	85%	84%

Table 35: Type-level evaluation of the LLM-based variation classifier on 49 manually annotated (preferred term, variant) pairs. Per-class precision and recall are reported; accuracy is a global metric (total correct / total) reported at the Total row only.

Results. The classifier achieves 84% overall type recall and 77% subtype accuracy across the 49 evaluable pairs (one pair was excluded as the variant was not a genuine terminological link). Performance is near-perfect for single-type variations: GRAPH, LEX, RED, EXP, and MORPH all reach 100% recall at the type level, with subtype accuracy above 85% for the two most frequent classes (GRAPH: 86%, LEX: 89%).

The only systematic failure concerns compound variations (COMBO): the classifier recovers only 1 out of 9 gold COMBO instances (recall 11%), con-

sistently predicting a single type instead of a combination. All 8 missed COMBO cases follow the same pattern, a multi-word term is abbreviated to an acronym that is also an ellipsis of the original (e.g., *topic detection / TDT, model / LM*), which the classifier labels as pure GRAPH (acronym) rather than the compound GRAPH+EXP that gold annotation assigns. This over-simplification depresses GRAPH precision to 67%, since seven COMBO pairs are absorbed into the GRAPH class.

The remaining subtype errors are minor: two GRAPH1 predictions should be GRAPH5 (partial abbreviation, e.g., *ASR system*), and two LEX1 predictions should be LEX2 (near-synonymy with head-word change, e.g., *similarity function, parallel text*).

Overall, the classifier is reliable for the large majority of variation types encountered in practice. The COMBO class is both the rarest in the sample and the hardest to predict, but its impact on downstream analyses is limited since compound variations are a small fraction of the full dataset.

Type	Support	Precision	Recall	Accuracy
GRAPH	14	57%	86%	—
MORPH	1	100%	100%	—
RED	4	100%	100%	—
EXP	2	100%	100%	—
LEX	19	85%	89%	—
COMBO	9	100%	11%	—
Total	49	90%	81%	76%

Table 36: Subtype-level evaluation of the LLM-based variation classifier on 49 manually annotated pairs, grouped by gold type. Precision is computed over all instances predicted as a given type; recall over gold instances of that type. Accuracy is global.

C Manual Evaluation of Alignment Accuracy

The alignment step (Section 3) links each source term occurrence to its target span using Bertalign and SimAlign, with an LLM fallback for low-confidence spans. To assess the reliability of these alignments, we conducted a manual evaluation on a sample of 100 term alignments drawn from the full set of pipeline outputs across all documents. To ensure balanced coverage, the sample was stratified across the nine system outputs, that is, the four MT systems under

both the baseline and base+terms conditions plus the human reference, each contributing eleven alignments (twelve for the reference); alignments with an empty target span were excluded. For each sampled alignment, one of the authors, a native French speaker fluent in English with expertise in computational linguistics, judged whether the aligned target span is the correct realization of the source term in the given sentence pair, and, when incorrect, recorded the error type. Table 38 reports accuracy per system and Table 39 lists the incorrect alignments.

System	Condition	#	Acc.
Llama	baseline	11	91%
Llama	base+terms	11	91%
Qwen	baseline	11	100%
Qwen	base+terms	11	100%
Euro9	baseline	11	100%
Euro9	base+terms	11	100%
Euro22	baseline	11	100%
Euro22	base+terms	11	91%
Ref	—	12	92%
Total		100	96%

Table 38: Per-system accuracy of the manual alignment evaluation on 100 alignments stratified across the nine system outputs. Each row reports the number of judged alignments and the proportion judged correct.

Results. The aligner is correct on 96% of the sampled alignments (96/100). Accuracy is high and comparable across conditions, at 98% for baseline (43/44), 95% for base+terms (42/44), and 92% for the human reference (11/12); the small per-system differences are not meaningful given that each system contributes only 11–12 judgements. Accuracy is likewise stable across variation types, with no error on the ten sampled dynamic variants.

The four errors fall into two categories. Three are lexical-choice errors, where the span captures an adjacent target word rather than the term itself: *corpus* aligned to *Récolte* (from *Corpus Harvesting / Récolte de corpus*), *indexing* to *discours* (from *speech indexing / indexation des discours*), and *future* echoed as *Future* where no distinct French realization appears in the sentence. The remaining case is a span-boundary error, where *automatic speech recog-*

nition was aligned to the truncated *reconnaissance de la parole* instead of *reconnaissance automatique de la parole*. All four affect single occurrences and none reflects a systematic failure mode, consistent with the low overall error rate.

D Prompts

This appendix collects the full prompts used by the LLM components of our pipeline: alignment disambiguation (Figure 2), baseline translation (Figure 3), glossary-guided translation (Figure 6), and variation classification (Figure 9).

E Reproducibility Statement

The variation classifier and the alignment-disambiguation step in our pipeline use OpenAI’s

gpt-4.1-mini (Section 3; prompts in Appendix D); this usage is intrinsic to the contribution and is fully described in the methodology. All MT inference was run on a single NVIDIA H100 96GB GPU. Total GPU time across 4 systems and 2 conditions (baseline and base+terms) for the 42 documents from ParaNLP and IWSLT2023 is approximately 10 minutes using vLLM (Kwon et al., 2023) with pre-generated prompts. The labeling pipeline issued approximately 5,000 requests to gpt-4.1-mini (snapshot gpt-4.1-mini-2026-04-01), for a total cost of approximately 15 USD. Reliance on a closed commercial LLM introduces a versioning risk: OpenAI silently retires model snapshots, so re-running the pipeline more than a few months from publication may result in slightly different labels.

```

"""You are a professional {source_language}-{target_language} translator, teaching the students the course
on technical translation. You are checking a student's translation of a sentence that contains a
technical term. You are given an {source_language} term (it can be a word or an expression), a source
{source_language} sentence containing this term (it may be cased differently or contain additional
punctuation), and a student's {target_language} translation. You need to find how the student has
translated the term in question in {target_language}, and return only that term.

Important: do not change the translated term anyhow, copy it straight from the sentence! For example, keep
the casing and the grammar form of the translated term as is.

When completing the task, follow the examples below:

{source_language} sentence: This paradigm based on ant colony algorithms for the exploration of the graph
removes the need to dynamically expand the graph: the memory footprint becomes independant of the
language model size.
{source_language} term: language model
{target_language} translation: Ce paradigme basé sur les algorithmes de colonie de fourmis pour l'
exploration du graphe supprime le besoin d'étendre dynamiquement le graphe : l'empreinte mémoire
devient indépendante de la taille du modèle de langage.
Translated term: modèle de langage
.
.
.
{source_language} sentence: We used our set of manually-written chronologies as a training corpus to
perform machine learning experiments.
{source_language} term: training corpus
{target_language} translation: Nous avons utilisé notre ensemble de chronologies rédigées manuellement
comme corpus d'apprentissage pour effectuer des expériences d'apprentissage automatique.
Translated term: corpus d'apprentissage

{source_language} sentence: {src_segment}
{source_language} term: {src_term}
{target_language} translation: {tgt_segment}
Translated term: """

```

Figure 2: Alignment prompt: maps a source term to its translation in the target sentence.

```

system prompt:
"""
You are a good translator! Translate the following text from English into French. Do not include any
extraneous note, commentary, explanations, or annotations. You must reply only with the translated
text in French.
"""
user prompt:
"""
English: $src\nFrench:
"""

```

Figure 3: Baseline translation prompt for Llama3.1-8B-Instruct.

```

system prompt:
'''
Translate the following English source text to French:
'''
user prompt:
'''
English: $src\nFrench:
'''

```

Figure 4: Baseline translation prompt for EuroLLM.

```

'''
You are a good translator! Translate the following text from English into French.\nEnglish: ${src}\nFrench:
'''

```

Figure 5: Baseline translation prompt for Qwen3-8B.

```

<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 26 Jul 2024

You are a good translator! Consider the provided terms in English and their French translations:
Terms: present = présent - borrowing = emprunt - corpus = corpus - annotated corpus = corpus annot  
Translate the following text from English into French. Do not include any extraneous note, commentary,
explanations, or annotations. You must reply only with the translated text in French.<|eot_id|><|
start_header_id|>user<|end_header_id|>

English: Hi, this is Elena and I'm going to be presenting our work, Detecting Unassimilated Borrowings in
Spanish: An Annotated Corpus and Approaches to Modeling.
French:<|eot_id|><|start_header_id|>assistant<|end_header_id|>

```

Figure 6: Term-injection translation prompt for Llama3.1-8B-Instruct.

```

<|im_start|>system
Translate the following English source text to French, considering the provided terms in English and their
French translations:
Terms: present = pr  sent - borrowing = emprunt - corpus = corpus - annotated corpus = corpus annot  <|im_end
|>
<|im_start|>user
English: Hi, this is Elena and I'm going to be presenting our work, Detecting Unassimilated Borrowings in
Spanish: An Annotated Corpus and Approaches to Modeling.
French:<|im_end|>
<|im_start|>assistant

```

Figure 7: Term-injection translation prompt for EuroLLM.

```

<|im_start|>user
You are a good translator! Consider the provided terms in English and their French translations:
Terms: present = présent - borrowing = emprunt - corpus = corpus - annotated corpus = corpus annoté
Translate the following text from English into French.
English: Hi, this is Elena and I'm going to be presenting our work, Detecting Unassimilated Borrowings in
        Spanish: An Annotated Corpus and Approaches to Modeling.
French: <|im_end|>
<|im_start|>assistant
<think>

</think>

```

Figure 8: Term-injection translation prompt for Qwen3-8B.

Source	We dealt with time-stamped events governed by a reported speech verb or otherwise appearing in reported speech .
Reference	Nous avons traité des événements datés régis par un verbe de parole rapportée ou apparaissant dans un discours rapporté .
Llama	Nous avons traité des événements datés régis par un verbe de discours rapporté ou apparaissant autrement dans le discours rapporté .
Euro22	Nous avons traité des événements marqués par un verbe de discours rapporté ou apparaissant autrement dans un discours rapporté .
Euro9	Nous avons traité des événements marqués par un horodatage régis par un verbe de discours rapporté ou apparaissant autrement dans un discours rapporté .
Qwen	Nous avons traité des événements datés de temps, régis par un verbe de discours direct ou autrement apparaissant dans un discours rapporté .

Table 15: A sentence from the ParaNLP corpus where the reference alternates between *parole rapportée* and *discours rapporté* for the same source term *reported speech*, both standard French renderings of the same concept. The four MT systems produce a single repeated form (Qwen substitutes *discours direct*, a different concept).

Source	Reference
... via Reinforcement Learning (RL).	... via l'apprentissage par renforcement (RL). <i>GRAPH</i> → <i>GRAPH</i> ✓
... a data-driven method such as RL .	... des méthodes d'apprentissage statistique telles que l'apprentissage par renforcement . <i>GRAPH</i> → <i>NO-VAR</i> ✗
An RL agent compares different ...	Un agent apprenant par renforcement compare différentes ... <i>GRAPH</i> → <i>COMBO</i> ✗

Table 16: Cross-term variation for *reinforcement learning* → *apprentissage par renforcement*: *RL* is preserved or expanded, illustrating type-dependent variability.

Table 18: Average proportion of terms covered by no evaluation method per document (*occ* = occurrences; *dist* = distinct terms; never evaluated $\equiv$ not in glossary $\wedge$ hapax ($k=1$) $\wedge$ not ITC-eligible)

Category	Avg occ/doc		Avg dist/doc	
	Avg <i>N</i>	%	Avg <i>N</i>	%
All terms	273.9	—	68.7	—
Eval. by ≥ 1 method	265.5	96.8%	60.3	87.9%
Never evaluated	8.4	3.2%	8.4	12.1%
Hapax ($k=1$, excl. by TLC)	34.2	12.5%	34.2	49.8%
Not in glossary (excl. by EM)	29.5	10.8%	6.7	9.8%

Averaged over 31 documents. Never evaluated $\equiv$ not in glossary $\wedge$ hapax $\wedge$ not ITC-eligible.

Category	Avg occ/doc		Avg dist/doc	
	Avg <i>N</i>	%	Avg <i>N</i>	%
All terms	120.3	—	35.2	—
Eval. by ≥ 1 method	117.3	97.5%	32.2	91.9%
Never evaluated	3.0	2.5%	3.0	8.1%
Hapax ($k=1$, excl. by TLC)	16.5	13.7%	16.5	46.9%
Not in glossary (excl. by EM)	7.4	6.2%	2.2	6.2%

Averaged over 10 documents. Never evaluated $\equiv$ not in glossary $\wedge$ hapax $\wedge$ not ITC-eligible.

Table 19: Average proportion of terms covered by no evaluation method per document (*occ* = occurrences; *dist* = distinct terms; never evaluated $\equiv$ not in glossary $\wedge$ hapax ($k=1$) $\wedge$ not ITC-eligible)

Var.	ParaNLP							IWSLT2023						
	<i>N</i>	%	Ref _S	Euro22	Euro9	Qwen	Llama	<i>N</i>	%	Ref	Euro22	Euro9	Qwen	Llama
NO-VAR	197.7	72.2%	85.31	83.15	83.77	81.54	81.80	88.4	73.5%	72.63	72.47	73.07	74.87	70.85
RED	24.6	9.0%	87.09	85.13	82.00	86.15	88.11	9.9	8.2%	67.92	65.42	65.42	68.75	66.67
MORPH	1.7	0.6%	0.00	0.00	0.00	0.00	0.00	0.3	0.2%	0.00	0.00	0.00	0.00	0.00
EXP	3.2	1.2%	6.45	6.45	23.12	2.15	23.12	0.7	0.6%	0.00	0.00	0.00	0.00	0.00
GRAPH	20.6	7.5%	45.76	43.11	42.51	45.08	42.00	11.4	9.5%	73.78	58.15	59.26	65.87	56.35
LEX	22.0	8.0%	40.35	47.79	49.63	38.59	45.24	7.4	6.2%	34.66	45.98	55.24	45.77	44.25
COMBO	4.0	1.5%	0.00	0.00	0.00	0.00	0.00	2.2	1.8%	0.00	0.00	0.00	0.00	0.00

Table 21: EM accuracy (%) by source variation category under the baseline condition. *N*: avg. occurrences per document; %: share of total occurrences. Best system per corpus in bold.

Var.	ParaNLP							IWSLT2023						
	<i>N</i>	%	Ref _S	Euro22	Euro9	Qwen	Llama	<i>N</i>	%	Ref	Euro22	Euro9	Qwen	Llama
NO-VAR	22.4	64.9%	88.96	90.82	90.67	88.20	90.11	13.2	70.6%	95.41	85.58	88.26	88.35	87.71
RED	3.8	11.0%	89.95	89.62	84.55	88.54	91.99	2.4	12.8%	94.24	84.86	82.94	82.38	87.44
MORPH	0.0	0.1%	1.61	1.61	1.61	1.61	1.61	0.0	0.0%	0.00	0.00	0.00	0.00	0.00
EXP	0.5	1.3%	26.87	22.65	25.84	18.62	24.48	0.1	0.5%	8.00	6.00	10.00	10.00	6.00
GRAPH	3.5	10.0%	55.89	65.18	67.74	65.19	67.68	1.3	7.0%	88.86	65.36	73.71	80.91	58.38
LEX	3.9	11.3%	87.66	78.66	85.06	81.45	84.19	1.3	7.0%	52.95	52.50	49.92	53.41	49.47
COMBO	0.5	1.3%	19.97	20.89	21.89	21.12	24.19	0.4	2.1%	19.17	12.50	12.50	14.17	14.17

Table 22: TCR pseudo-ref (%) by source variation category under the baseline condition. *N*: avg. distinct terms per document with $k \geq 2$ occurrences; %: share of total. Best system per corpus in bold.

Var.	ParaNLP							IWSLT2023						
	<i>N</i>	%	Ref _S	Euro22	Euro9	Qwen	Llama	<i>N</i>	%	Ref	Euro22	Euro9	Qwen	Llama
NO-VAR	22.4	64.9%	87.98	90.79	89.69	87.50	88.71	13.2	70.6%	94.81	86.60	86.63	85.78	84.85
RED	3.8	11.0%	84.97	86.31	82.64	84.23	89.23	2.4	12.8%	89.43	72.43	72.29	70.84	80.71
MORPH	0.0	0.1%	0.00	0.00	0.00	0.00	0.00	0.0	0.0%	0.00	0.00	0.00	0.00	0.00
EXP	0.5	1.3%	24.20	18.00	20.83	13.86	24.13	0.1	0.5%	5.00	0.00	10.00	10.00	5.00
GRAPH	3.5	10.0%	54.49	60.30	63.54	64.16	67.23	1.3	7.0%	88.24	55.36	61.23	73.32	53.82
LEX	3.9	11.3%	84.00	74.87	81.65	78.19	77.61	1.3	7.0%	51.73	47.07	46.61	49.64	44.53
COMBO	0.5	1.3%	13.30	14.29	17.19	15.44	19.88	0.4	2.1%	17.78	6.67	6.67	8.89	8.89

Table 23: TCR temporal (%) by source variation category under the baseline condition. *N*: avg. distinct terms per document with $k \geq 2$ occurrences; %: share of total. Best system per corpus in bold.

Var.	ParaNLP							IWSLT2023						
	<i>N</i>	%	Ref _S	Euro22	Euro9	Qwen	Llama	<i>N</i>	%	Ref	Euro22	Euro9	Qwen	Llama
NO-VAR	22.4	80.7%	86.57	84.57	85.17	82.95	83.33	13.2	80.0%	74.01	73.92	74.05	76.70	72.28
RED	1.4	5.1%	87.82	86.36	83.19	88.58	89.21	1.2	7.3%	72.08	69.58	69.58	75.42	70.83
MORPH	0.0	0.1%	0.00	0.00	0.00	0.00	0.00	0.0	0.0%	0.00	0.00	0.00	0.00	0.00
EXP	0.2	0.7%	1.61	1.61	7.53	0.54	7.53	0.1	0.6%	0.00	0.00	0.00	0.00	0.00
GRAPH	2.3	8.1%	44.78	41.81	41.06	44.45	40.71	1.3	7.9%	76.61	60.36	61.46	68.66	58.49
LEX	1.4	5.0%	37.90	46.82	47.36	38.14	42.47	0.7	4.2%	13.89	25.35	25.03	25.00	23.76
COMBO	0.1	0.2%	0.00	0.00	0.00	0.00	0.00	0.0	0.0%	0.00	0.00	0.00	0.00	0.00

Table 24: TCR_{EM_{macro}} (%) by source variation category under the baseline condition. *N*: avg. in-glossary distinct terms per document with $k \geq 2$ occurrences; %: share of total. Best system per corpus in bold.

Var.	ParaNLP							IWSLT2023						
	<i>N</i>	%	Ref _S	Euro22	Euro9	Qwen	Llama	<i>N</i>	%	Ref	Euro22	Euro9	Qwen	Llama
GRAPH	9.1	34.2%	57.95	69.92	72.06	67.73	77.83	1.0	33.3%	90.00	85.71	85.71	50.00	85.71
RED	8.4	31.5%	87.36	84.85	81.03	84.65	83.95	1.1	36.7%	81.82	95.65	100.00	100.00	100.00
EXP	1.3	4.8%	40.00	32.43	29.73	23.68	28.21	0.2	6.7%	100.00	100.00	50.00	0.00	0.00
LEX	7.9	29.5%	56.97	70.35	72.12	70.76	69.71	0.7	23.3%	85.71	60.00	60.00	50.00	60.00

Table 25: CTV variation preservation (%) by source variation category under the baseline condition. *N*: avg. eligible occurrences per document (variants whose head term is translated as NO-VAR); %: share of eligible total. Best system per corpus in bold.

Var.	ParaNLP							IWSLT2023						
	<i>N</i>	%	Ref _S	Euro22	Euro9	Qwen	Llama	<i>N</i>	%	Ref	Euro22	Euro9	Qwen	Llama
NO-VAR	197.7	72.2%	85.31	93.23	90.38	93.96	93.63	88.4	73.5%	72.63	87.12	80.69	90.22	89.31
RED	24.6	9.0%	87.09	88.75	86.14	89.39	89.56	9.9	8.2%	67.92	70.00	65.42	72.50	72.50
MORPH	1.7	0.6%	0.00	0.00	0.00	0.00	0.00	0.3	0.2%	0.00	0.00	0.00	0.00	0.00
EXP	3.2	1.2%	6.45	12.90	12.90	24.73	22.58	0.7	0.6%	0.00	0.00	0.00	0.00	0.00
GRAPH	20.6	7.5%	45.76	44.40	40.98	43.41	48.61	11.4	9.5%	73.78	63.41	58.57	78.41	76.27
LEX	22.0	8.0%	40.35	47.31	48.12	37.01	44.01	7.4	6.2%	34.66	46.54	48.08	47.71	42.59
COMBO	4.0	1.5%	0.00	0.00	0.00	0.00	0.00	2.2	1.8%	0.00	0.00	0.00	0.00	0.00

Table 26: EM accuracy (%) by source variation category under the base+terms condition. *N*: avg. occurrences per document; %: share of total occurrences. Best system per corpus in bold.

Var.	ParaNLP							IWSLT2023						
	<i>N</i>	%	Ref _S	Euro22	Euro9	Qwen	Llama	<i>N</i>	%	Ref	Euro22	Euro9	Qwen	Llama
NO-VAR	22.4	64.9%	88.96	95.28	93.22	94.98	94.01	13.2	70.6%	95.41	90.70	91.49	94.98	93.47
RED	3.8	11.0%	89.95	91.14	87.83	91.61	92.15	2.4	12.8%	94.24	82.44	85.44	80.89	86.06
MORPH	0.0	0.1%	1.61	1.61	1.61	1.61	1.61	0.0	0.0%	0.00	0.00	0.00	0.00	0.00
EXP	0.5	1.3%	26.87	22.61	25.69	23.46	24.01	0.1	0.5%	8.00	10.00	10.00	6.00	6.00
GRAPH	3.5	10.0%	55.89	58.55	60.92	59.85	61.30	1.3	7.0%	88.86	74.46	73.60	81.46	79.89
LEX	3.9	11.3%	87.66	85.32	88.17	79.56	80.29	1.3	7.0%	52.95	54.32	52.65	49.32	48.11
COMBO	0.5	1.3%	19.97	21.43	23.35	22.81	23.66	0.4	2.1%	19.17	12.50	12.50	12.50	13.33

Table 27: TCR pseudo-ref (%) by source variation category under the base+terms condition. *N*: avg. distinct terms per document with $k \geq 2$ occurrences; %: share of total. Best system per corpus in bold.

Var.	ParaNLP							IWSLT2023						
	<i>N</i>	%	Ref _S	Euro22	Euro9	Qwen	Llama	<i>N</i>	%	Ref	Euro22	Euro9	Qwen	Llama
NO-VAR	22.4	64.9%	87.98	94.87	92.76	95.10	93.72	13.2	70.6%	94.81	89.86	91.53	94.15	92.52
RED	3.8	11.0%	84.97	88.03	87.68	88.56	89.14	2.4	12.8%	89.43	75.71	73.95	71.24	82.05
MORPH	0.0	0.1%	0.00	0.00	0.00	0.00	0.00	0.0	0.0%	0.00	0.00	0.00	0.00	0.00
EXP	0.5	1.3%	24.20	20.45	23.71	21.89	18.36	0.1	0.5%	5.00	10.00	10.00	0.00	5.00
GRAPH	3.5	10.0%	54.49	54.51	54.78	56.08	56.27	1.3	7.0%	88.24	63.51	61.22	71.10	69.85
LEX	3.9	11.3%	84.00	81.54	86.06	74.26	77.26	1.3	7.0%	51.73	49.42	48.51	46.29	40.64
COMBO	0.5	1.3%	13.30	17.19	17.86	17.36	18.26	0.4	2.1%	17.78	7.78	8.89	6.67	8.89

Table 28: TCR temporal (%) by source variation category under the base+terms condition. *N*: avg. distinct terms per document with $k \geq 2$ occurrences; %: share of total. Best system per corpus in bold.

Var.	ParaNLP							IWSLT2023						
	<i>N</i>	%	Ref _S	Euro22	Euro9	Qwen	Llama	<i>N</i>	%	Ref	Euro22	Euro9	Qwen	Llama
NO-VAR	22.4	80.7%	86.57	94.15	91.32	94.59	94.38	13.2	80.0%	74.01	88.09	81.32	91.39	89.68
RED	1.4	5.1%	87.82	89.80	86.90	90.41	90.24	1.2	7.3%	72.08	74.17	69.58	76.67	76.67
MORPH	0.0	0.1%	0.00	0.00	0.00	0.00	0.00	0.0	0.0%	0.00	0.00	0.00	0.00	0.00
EXP	0.2	0.7%	1.61	1.61	1.61	5.91	6.99	0.1	0.6%	0.00	0.00	0.00	0.00	0.00
GRAPH	2.3	8.1%	44.78	43.58	39.93	43.16	48.63	1.3	7.9%	76.61	66.21	61.35	71.21	67.64
LEX	1.4	5.0%	37.90	46.66	45.74	33.94	39.35	0.7	4.2%	13.89	25.91	27.73	16.94	24.60
COMBO	0.1	0.2%	0.00	0.00	0.00	0.00	0.00	0.0	0.0%	0.00	0.00	0.00	0.00	0.00

Table 29: TCR_{EM_{macro}} (%) by source variation category under the base+terms condition. *N*: avg. in-glossary distinct terms per document with $k \geq 2$ occurrences; %: share of total. Best system per corpus in bold.

Var.	ParaNLP							IWSLT2023						
	<i>N</i>	%	Ref _S	Euro22	Euro9	Qwen	Llama	<i>N</i>	%	Ref	Euro22	Euro9	Qwen	Llama
GRAPH	9.1	34.2%	57.95	60.23	65.47	63.28	66.22	1.0	33.3%	90.00	85.71	85.71	90.00	90.00
RED	8.4	31.5%	87.36	82.98	80.29	83.57	81.42	1.1	36.7%	81.82	83.33	95.65	85.71	86.67
EXP	1.3	4.8%	40.00	37.25	32.00	62.50	36.54	0.2	6.7%	100.00	50.00	0.00	100.00	100.00
LEX	7.9	29.5%	56.97	64.23	66.93	63.24	62.55	0.7	23.3%	85.71	60.00	60.00	60.00	60.00

Table 30: CTV variation preservation (%) by source variation category under the base+terms condition. *N*: avg. eligible occurrences per document (variants whose head term is translated as NO-VAR); %: share of eligible total. Best system per corpus in bold.

Type	Subtype	Description and Examples
<i>Single-Type Variations</i>		
GRAPH	GRAPH 1	Acronyms/Initialisms (e.g., <i>World Health Organization</i> / <i>WHO</i>)
	GRAPH 2	Symbols/Formulas (e.g., <i>carbon dioxide</i> / <i>CO₂</i>)
	GRAPH 3	Spelling changes (e.g., <i>micro-organism</i> / <i>micro organism</i>)
	GRAPH 4	Multiple spelling variants (e.g., <i>micro-organism</i> / <i>microorganism</i>)
	GRAPH 5	Partial abbreviation (e.g., <i>chemotherapy</i> / <i>chemo</i>)
MORPH	MORPH 1	Word order (e.g., <i>automatic translation system</i> / <i>translation system automatic</i>)
	MORPH 2	Article presence/absence (e.g., <i>collection cost</i> / <i>cost of the collection</i>)
	MORPH 3	Inflection (e.g., <i>adaptation</i> / <i>adaptations</i>)
	MORPH 4	Derivation/Affixation (e.g., <i>white</i> / <i>whiten</i>)
	MORPH 5	Syntactic structure (e.g., <i>carbon emission</i> / <i>emission of carbon</i>)
RED	RED 1	Base reduction (e.g., <i>automatic translation system</i> / <i>translation system</i>)
	RED 2	Extension reduction (e.g., <i>automatic translation system</i> / <i>automatic system</i>)
	RED 3	Other reductions (e.g., <i>medium-sized enterprise</i> / <i>enterprise</i>)
EXP	EXP 1	Semantic addition (e.g., <i>translation system</i> / <i>automatic translation system</i>)
	EXP 2	Explicit form (e.g., <i>stats</i> / <i>statistics</i>)
	EXP 3	Lexical insertion (e.g., <i>dark urine</i> / <i>dark colored urine</i>)
	EXP 4	Abbreviation expansion (e.g., <i>ex.</i> / <i>example</i>)
LEX	LEX 1	Simple unit substitution (e.g., <i>residues</i> / <i>waste</i>)
	LEX 2	Base change (e.g., <i>translation system</i> / <i>translation model</i>)
	LEX 3	Extension change (e.g., <i>binary classification model</i> / <i>logistic regression model</i>)
	LEX 4	Base and extension change (e.g., <i>coronavirus disease 2019</i> / <i>Wuhan pneumonia</i>)
<i>Multiple-Type Variations</i>		
CM	GRAPH +LEX	Graphical and lexical (e.g., <i>alpha particle</i> / <i>α ray</i>)
	LEX +RED	Lexical and reduction (e.g., <i>translation system</i> / <i>model</i>)
	GRAPH +RED	Graphical and reduction (e.g., <i>micro-organism pathogen</i> / <i>microorganism</i>)
	GRAPH +MORPH	Graphical and morphosyntactic (e.g., <i>micro-organisms</i> / <i>microorganism</i>)
	MORPH +LEX	Morphosyntactic and lexical (e.g., <i>system of translation</i> / <i>translation model</i>)
	EXP +LEX	Expansion and lexical (e.g., <i>model</i> / <i>automatic translation system</i>)
	Other	Other combinations of two or more types (format: CODE1+CODE2+...)

Table 34: Terminological variation typology with subtypes and examples. GRAPH: Graphical, MORPH: Morphosyntactic, RED: Reduction, EXP: Expansion, LEX: Lexical, CM: Combined Multiple.

#	Head term	Variant	Pred. type	Pred. subtype	Gold type	Gold subtype
1	word error rate	WER	GRAPH	GRAPH 1	GRAPH	GRAPH 1
2	hidden Markov model	HMM	GRAPH	GRAPH 1	GRAPH	GRAPH 1
3	automatic speech recognition system	ASR system	GRAPH	GRAPH 1	GRAPH	GRAPH 5
4	data set	dataset	GRAPH	GRAPH 3	GRAPH	GRAPH 3
5	tokenization	tokenisation	GRAPH	GRAPH 3	GRAPH	GRAPH 3
6	named entity recognition	NER	GRAPH	GRAPH 1	GRAPH	GRAPH 1
7	topic detection	TDT	GRAPH	GRAPH 1	COMBO	GRAPH 1 + EXP 1
8	conditional random field	CRF	GRAPH	GRAPH 1	GRAPH	GRAPH 1
9	Vector space model	VSM	GRAPH	GRAPH 1	GRAPH	GRAPH 1
10	optimization	optimisation	GRAPH	GRAPH 3	GRAPH	GRAPH 3
11	phrase	PP	GRAPH	GRAPH 1	COMBO	GRAPH 1 + EXP 1
12	information retrieval model	IR model	GRAPH	GRAPH 1	GRAPH	GRAPH 5
13	cluster	IC	GRAPH	GRAPH 1	COMBO	GRAPH 1 + EXP 1
14	reinforcement learning	RL	GRAPH	GRAPH 1	GRAPH	GRAPH 1
15	context-free grammar	CFG	GRAPH	GRAPH 1	GRAPH	GRAPH 1
16	Bayesian network	DBN	GRAPH	GRAPH 1	COMBO	GRAPH 1 + EXP 1
17	n-gram	n gram	GRAPH	GRAPH 3	GRAPH	GRAPH 3
18	speech recognition	TTS	GRAPH	GRAPH 1	NOT LINK	
19	finite state automata	DFA	GRAPH	GRAPH 1	COMBO	GRAPH 1 + EXP 1
20	syntactic category	POS	GRAPH	GRAPH 1	COMBO	GRAPH 1 + LEX 4
21	model	LM	GRAPH	GRAPH 1	COMBO	GRAPH 1 + EXP 1
22	self-organizing map	SOM	GRAPH	GRAPH 1	GRAPH	GRAPH 1
23	input representation	input sequence	LEX	LEX 3	LEX	LEX 3
24	structural annotation	segmentation	LEX	LEX 1	LEX	LEX 1
25	validation data	development data	LEX	LEX 3	LEX	LEX 3
26	elision	deletion	LEX	LEX 1	LEX	LEX 1
27	inference	entailment	LEX	LEX 1	LEX	LEX 1
28	phrase	VP	LEX	LEX 1	COMBO	GRAPH 1 + EXP 1
29	similarity measure	similarity function	LEX	LEX 1	LEX	LEX 2
30	parallel corpus	parallel text	LEX	LEX 1	LEX	LEX 2
31	syntactic category	word class	LEX	LEX 1	LEX	LEX 1
32	parser	analyzer	LEX	LEX 1	LEX	LEX 1
33	lexeme	lexical item	LEX	LEX 1	LEX	LEX 1
34	interjection	exclamation	LEX	LEX 1	LEX	LEX 1
35	syntactic transfer	reordering	LEX	LEX 1	LEX	LEX 1
36	alignment	matching	LEX	LEX 1	LEX	LEX 1
37	true negative rate	specificity	LEX	LEX 1	LEX	LEX 1
38	Markov model	Markov chain	LEX	LEX 2	LEX	LEX 2
39	syntactic category	part-of-speech	LEX	LEX 1	LEX	LEX 1
40	training data	training dataset	LEX	LEX 2	LEX	LEX 2
41	foundation model	base model	LEX	LEX 3	LEX	LEX 3
42	lexeme	lexical unit	LEX	LEX 1	LEX	LEX 1
43	multi-word term	multi-word	RED	RED 1	RED	RED 1
44	grammatical subject	subject	RED	RED 2	RED	RED 2
45	programming language C++	C++	RED	RED 1	RED	RED 1
46	Apache OpenNLP	OpenNLP	RED	RED 1	RED	RED 1
47	speech recognition	automatic speech recognition	EXP	EXP 1	EXP	EXP 1
48	future	future tense	EXP	EXP 3	EXP	EXP 3
49	syntactic parsing	syntactical analysis	COMBO	LEX 3+MORPH 4	COMBO	LEX 3+MORPH 4
50	similarity measure	measure of similarity	MORPH	MORPH 5	MORPH	MORPH 5

Table 37: The 50 manually evaluated (preferred term, variant) pairs with predicted and gold variation labels.

Incorrect alignment in context (source term and aligned target span in bold)	
Llama (baseline), lexical choice: <i>corpus</i> → <i>Récolte</i> (should align <i>corpus</i>)	
EN	Corpus Harvesting
FR	Récolte de corpus
Llama (base+terms), lexical choice: <i>future</i> → <i>Future</i> (source form echoed; no distinct French realization)	
EN	Conclusion and Future Work
FR	Conclusion et Travail à Faire
Euro22 (base+terms), lexical choice: <i>indexing</i> → <i>discours</i> (should align <i>indexation</i>)	
EN	speech indexing
FR	indexation des discours
Ref, span boundary: <i>automatic speech recognition</i> → <i>reconnaissance de la parole</i> (drops <i>automatique</i>)	
EN	The work presented here aims at expanding this paradigm to automatic speech recognition .
FR	Les travaux présentés ici visent à étendre ce paradigme à la reconnaissance automatique de la parole .

Table 39: The four incorrect alignments among the 100 manually evaluated pairs, shown in source (EN) and target (FR) context. The source term is bold in the English segment and the aligned target span is bold in the French segment; for the span-boundary error, the non-bold word (*automatique*) is the token wrongly excluded from the span.

""You are an expert in scientific terminology and linguistic variation, teaching a course on terminology. A student has identified a potential variant of a technical term and needs your help to classify the relationship between the head term and its variant according to the established typology.

You receive:

- A head term (the preferred form of a technical term)
- A variant (an alternative form that may or may not be related to the head term)
- Context examples showing how these terms appear in real scientific texts (when available)

Your task: Determine the type of variation between the head term and the variant. Classify according to the variation typology. Return ONLY a JSON object with the exact format shown in the examples below.

Important rules:

- 1) First verify that the variant is actually related to the head term
- 2) Count the TYPES of changes, not the number of words: one type -> \VG/\VMS/\VR/\VE/\VL, multiple types -> CM
- 3) For CM (Multiple Changes), combine codes with + (e.g., "\VG+\VL" or "\VMS+\VR+\VL")
- 4) Copy the exact JSON format from the examples - no markdown, no extra text
- 5) Keep justifications brief (1-2 sentences maximum)

Reference of variation types:

\VG (Graphical): Written form changes without semantic change (acronyms, spelling, symbols)

- > \VG1: Acronym (European Parliament/EP), \VG2: Symbol (CO2), \VG3: Spelling (trade mark/trademark), \VG4: Multiple spelling, \VG5: Partial abbreviation

\VMS (Morphosyntactic): Structure or inflection changes (word order, articles, inflection, derivation)

- > \VMS1: Order, \VMS2: Article, \VMS3: Inflection, \VMS4: Derivation, \VMS5: Structure

\VR (Reduction): Element deletion

- > \VR1: Base reduction, \VR2: Extension reduction, \VR3: Other reduction

\VE (Expansion): Element addition

- > \VE1: Semantic addition, \VE2: Explicit form, \VE3: Lexical insertion, \VE4: Abbreviation development

\VL (Lexical): Lexical substitution (synonyms, near-synonyms, translations validated as variant in context)

- > \VL1: Simple unit, \VL2: Base change, \VL3: Attribute/extension change, \VL4: Base+extension change

CM (Multiple Changes): Combination of two or more types above

When completing the task, follow the examples below:

Head term: European Parliament
Variant: EP
{{"categorie": "\VG", "sous_type": "\VG1", "justification": "Acronym formed from the initials of the head term."}}

.

.

.

Head term: coronavirus disease 2019
Variant: COVID-19
{{"categorie": "CM", "sous_type": "\VG+\VR", "justification": "Acronym formation (\VG1) combined with structural reduction."}}

Head term: airborne dust particle
Variant: airborne dust
{{"categorie": "\VR", "sous_type": "\VR1", "justification": "Base reduction: deletion of 'particle' from the base noun phrase."}}

{context_section}

Head term: {head_term}
Variant: {variant}""

Figure 9: Variation classification prompt.

Layout-Based Chunk Alignment: Utilizing Visual Information to Collect Parallel Texts From Image Documents

Masaki Kinouchi
Kayoko Nohara
Xinru Zhu
Yuma Miura

School of Environment and Society, Institute of Science Tokyo, Tokyo, Japan

kinouchi.m.a7be@m.isct.ac.jp
nohara.k.d485@m.isct.ac.jp
zhu.x.1137@m.isct.ac.jp
tlfwu078@gmail.com

Abstract

This study proposes a layout-based chunk alignment method (Layout-CA) as an intermediate step between document- and sentence-level parallel text alignment for bilingual document images. Visually rich printed materials, such as institutional reports and magazines, often contain high-quality translations and are valuable sources of parallel data, yet their layout cues are underutilized. Layout-CA aligns semantically coherent text chunks across document pairs by integrating multi-modal cues from textual content and layout, and sentence alignment is then performed within the aligned chunk pairs. Experiments on English UNESCO reports and their Japanese translations show that introducing chunk alignment improves downstream sentence alignment for both Bleualign and Vecalign. When document order is disrupted, Layout-CA preserves alignment coverage by restricting sentence matching to corresponding chunks, enabling robust alignment in multilingual image documents.

1 Introduction

The construction of high-quality parallel corpora has long been a foundational topic in machine translation. Recent advances in corpus construction, driven by the vast availability of web documents, have enabled the collection of large-scale parallel data (Bañón et al., 2020; Morishita et al., 2020; Koehn, 2024), but ensuring their quality still remains a challenge. A key component in building parallel corpora is accurate sentence alignment, which is not only essential for the development of robust machine translation systems in general, but also plays a critical role in other language research fields, such as corpus linguistics and translation studies (Baker, 1993, 1996; Baorong, 2021). In these fields, it is often required to extract professional translations from printed documents to guarantee the quality of product in target. Although many alignment methods, especially at sentence level, have al-

ready been proposed, ranging from statistical (Gale and Church, 1993; Varga et al., 2007) or machine-translation-based approaches (Sennrich and Volk, 2011) to embedding-based approaches (Thompson and Koehn, 2019; Liu and Zhu, 2022), there is less attention to the complexities inherent in multilingual digitized documents or PDFs.

In practice, many professionally translated materials, such as institutional reports, magazines, etc., are published in visually rich layouts. These documents are often of high translation quality, making them valuable sources for parallel data. They contain rich visual information, such as complex layouts, font size, color and elements such as tables, figures, and captions that can indicate structural boundaries and correspondences. Conventional methods, which treat documents as simple sequences of text, leave room for further development to handle such structurally complex data when deal-

ing with documents as images.

To better capture the alignment structure of these documents, we draw inspiration from how humans approach translation alignment. When attempting to locate the translation of a particular sentence, one does not search the entire document linearly. Instead, one first identifies the corresponding article or section (based on headings, layout, or other visual cues), narrows down to a relevant paragraph, and finally locates the translated sentence. This top-down search reflects a hierarchical alignment process moving from document-level structure to semantically coherent chunks, down to sentences.

Motivated by this observation, we argue for an intermediate alignment step that operates at the level of chunk. Following the context of previous research (Zhao et al., 2025; Duarte et al., 2024), a *chunk* in this research is defined as a region for coherent texts in which the reader senses a persistent narrative theme while reading a document. These are semantically meaningful textual units such as articles, sections, or grouped paragraphs. These chunks can often be visually identified by layout features, such as headers indicated by font size or bounding boxes. In this paper, we propose a chunk-level text alignment framework for paired bilingual document images. By integrating layout-aware chunking into the pipeline, this work presents one of the first attempts to systematically incorporate multi-modal information into bilingual document alignment. Starting from document images, the framework applies OCR and document layout analysis to recover text content and structural information. It then uses them to align semantically corresponding chunks across document pairs before sentence-level alignment.

2 Related Work

2.1 Sentence Alignment

Sentence-level bilingual alignment research has evolved from simple statistical techniques to sophisticated neural methods, while leaving difficulties in extracting sentence pairs from raw image documents with text-based aligners. Early approaches like Gale-Church’s algorithm (Gale and Church, 1993) align sentences by maximizing length-based probability models via dynamic programming. Later algorithms combined sentence-length models with bilingual word correspondences

to improve alignment accuracy (Moore, 2002). Hybrid systems integrated multiple signals; for example, Hunalign (Varga et al., 2007) uses dictionaries plus length heuristics, and Bleualign (Sennrich and Volk, 2011) leverages rough machine translations with BLEU-based semantic similarity scoring to anchor matching sentences. In recent years, neural and embedding-based methods recast alignment as a bilingual retrieval problem: sentences are embedded into a shared multilingual vector space, and parallel pairs are identified by high semantic similarity (Thompson and Koehn, 2019; Liu and Zhu, 2022). These advances enable mining parallel sentences from documents for corpus construction. However, purely text-based aligners struggle with extracting from raw image documents as texts must be recovered from visually structured pages. Earlier work on parallel Web pages showed that HTML structure and text-chunk lengths can support translation detection and preliminary alignment (Resnik and Smith, 2003). This suggests that visual structure may likewise provide a valuable alignment signal in image documents, motivating chunk-aware approaches that exploit such information.

2.2 Document Layout Analysis

Document layout analysis (DLA) has been advanced by developments of CNN-based region detectors (Hao et al., 2016; Schreiber et al., 2017; Yang et al., 2017; Soto and Yoo, 2019). Recent work such as LayoutLM (Xu et al., 2020) introduces an architecture that integrates text, 2D layout coordinates, and visual features into multi-modal embedding, enabling strong performance on layout-aware understanding tasks. Building on this line, models such as LayoutReader (Wang et al., 2021) frame layout analysis as a reading-order recovery problem, explicitly modeling document structure through sequence prediction over detected regions.

Progress in DLA has been driven by large-scale annotated datasets. PubLayNet (Zhong et al., 2019) provides hundreds of thousands of document pages with automatically derived layout annotations, facilitating robust pretraining for region detection. DocLayNet (Pfitzmann et al., 2022) complements this with high-quality human annotations across diverse document types and finer-grained classes, improving generalization to real-world layouts. In addition, datasets such as ReadingBank (Wang et al.,

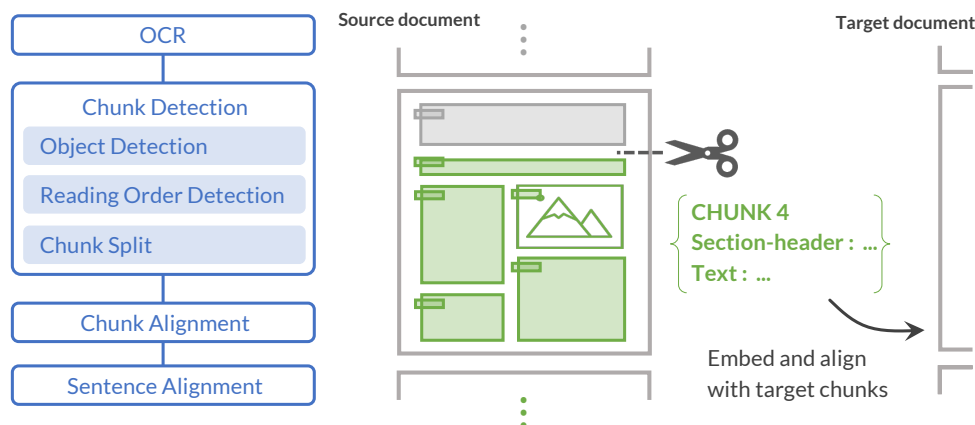


Figure 1: A diagram of layout-based chunk alignment method (Layout-CA).

2021) introduce explicit reading-order annotations, supporting the training and evaluation of models that reason over document flow rather than isolated regions.

3 Method

In this study, we propose layout-based chunk alignment method (Layout-CA), which consists of two principal stages (Figure 1). In the first stage, chunk detection, we transform raw document images into a linearized sequence of semantically coherent units, or chunks. This involves combining object detection with OCR, recovering reading order in complex multi-column layouts, and splitting the resulting sequence into segments that reflect the underlying discourse structure. In the second stage, chunk alignment, we establish bilingual correspondences between source and target chunks by computing similarity scores. The chunk layer functions as an intermediate alignment step that reduces the search space for sentence alignment while anchoring correspondences in document structure.

3.1 Chunk Detection

The task of chunk detection is to recover units of text that mirror the way readers and editors naturally perceive documents. Simply applying OCR to a page yields a collection of tokens or lines without structural coherence; on the other hand, treating detected layout regions as fixed units ignores the logical flow of reading. Our approach resolves this by integrat-

ing three sub-steps: OCR + object detection, reading order detection, and chunk splitting.

3.1.1 OCR + Object Detection

We first apply OCR and object detection page by page to capture the coordinates and labels of text boxes and textual information inside. We adopt labels defined in DocLayNet (Pfitzmann et al., 2022) and label each box using 11 categories: *Title*, *Section-header*, *Text*, *List-item*, *Formula*, *Picture*, *Table*, *Caption*, *Footnote*, *Page-header*, and *Page-footer*. These reflect common typographic and paratextual conventions of professionally edited publications. The OCR system outputs word-level tokens with bounding boxes, which are assigned to the detected layout regions based on maximum overlap. In this way, each region is enriched with its textual content, enabling joint reasoning over visual and linguistic signals. We also compute for each page a reference text height $\bar{h}_{Text}$ defined as the mean token height across *Text* segments. This statistic provides a robust scale for chunk split by identifying visually salient headers, since titles and section headers generally have larger font sizes than body text.

3.1.2 Reading Order Detection

We apply reading order detection to arrange text boxes linearly. Image documents often store multi-column layouts, sidebars, or figures, which break the

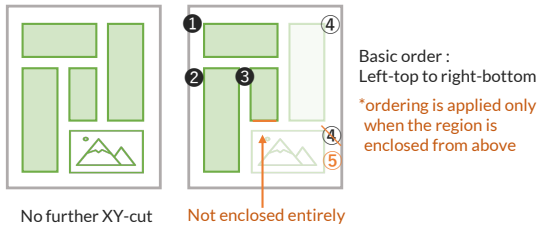


Figure 2: Reading order detection algorithm

simple top-to-bottom, left-to-right assumption. A key step, therefore, is to recover a plausible reading order that approximates how human readers traverse the page. We apply an XY-cut (George and Seth, 1984; Ha et al., 1995), which recursively identifies horizontal and vertical gaps in the distribution of bounding boxes and partitions the page into regions that correspond to columns or blocks.

However, as Liu et al. (2025) points out, XY-cut has a limitation in case like Figure 2. A further refinement prevents erroneous jumps across columns. Specifically, when linking a segment a to a next candidate segment b below it, we require that the horizontal coverage of above segments is greater than the width of a . This constraint enforces local column continuity. Finally, we normalize the position of paratextual segments: *Page-header* and *Page footer* are relocated to the beginning and end of the sequence in a page respectively.

3.1.3 Chunk Split

Having obtained a reading order, we next group segments into larger units that we call chunks. The central observation is that documents are typically organized around visually salient headers, such as article titles or section headers, which serve as anchors for subsequent blocks of text. To formalize this, we define salience r as a relative height ratio.

$$r = \bar{h} / \bar{h}_{Text}$$

$\bar{h}$ is the average height of tokens in the segment and $\bar{h}_{Text}$ is the reference average height of tokens labeled *Text* in the page. A new chunk is initiated when a segment is labeled as *Title* or *Section-header* and has salience r than or equal to a threshold r_{th} . Once a head is detected, all subsequent segments in reading order are appended to the same chunk until the next head is encountered. In this way, the

document is decomposed into a sequence of semantically coherent, layout-aware chunks that reflect its discourse organization.

3.2 Chunk Alignment

The task of chunk alignment aims to establish correspondences between chunks across source and target documents. Let c_i^s and c_j^t denote the sets of source- and target-language chunks in reading order. We calculate a similarity score that integrates semantic and layout features to identify pairs of chunks between the languages.

Each chunk is embedded using a multilingual sentence encoder. In our implementation, we adopt multilingual Sentence-BERT¹ (Reimers and Gurevych, 2019), which provides language-agnostic embeddings in a shared multilingual semantic space. We embed the texts label by label in each chunk with mean pooling. Then we calculate cosine similarity s_l between the same-label texts in two chunks and normalize it with sigmoid function. Empty fields are mapped to the zero vector.

For each candidate chunk pair (i, j) , we calculate a simple score function with s_l , label weight w_l and bias b :

$$S(i, j) = \sum_l w_l \cdot \sigma(s_l) + b$$

We trained a set of weights with a logistic regression model on annotated development dataset. The final alignment retains only pairs whose scores exceed the threshold. A source chunk is allowed to align with multiple target chunks (1:n), and the same target chunk can be shared across alignments with different source chunks.

3.3 From Chunks to Sentences

Layout-CA is integrated with existing sentence alignment methods by performing sentence alignment independently within each aligned chunk pair. The ultimate objective of our pipeline is to obtain reliable sentence-level bilingual alignments from raw image documents. The chunk layer contributes to this by constraining the alignment search space: rather than attempting to align sentences across entire documents, it performs within the narrower windows defined by aligned chunks.

¹<https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2>

ID	Type	Pages	Chunks	Sentences	OCR
UNESCO (en)	5 documents	50	18	1425	Tesseract
UNESCO (ja)	5 documents, same order as (en)	51	18	1463	Yomitoku
UNESCO (shuffled, ja)	5 documents, randomly shuffled	51	18	1463	Yomitoku

Table 1: Evaluation dataset

4 Experiments

We conducted experiments on English UNESCO reports and their Japanese translations to evaluate chunk detection, chunk alignment, and their impact on improving sentence alignment accuracy. For OCR, we use Tesseract OCR² for English and Yomitoku³ for Japanese. For object detection, we use YOLOv12 (Tian et al., 2025) pretrained on DocLayNet, while for Japanese, we further fine-tuned a model using the Japanese documents in the data. As a baseline for sentence alignment, we use the widely used Bleualign (Sennrich and Volk, 2011) with NLLB⁴ (Team et al., 2022) as the MT system and Vecalign (Thompson and Koehn, 2019) with Sentence-BERT as a multilingual embedder. The threshold of the chunk similarity score was empirically set to 0.3 based on preliminary experiments.

4.1 Evaluation Dataset

We evaluate on documents collected from UNESCO Digital Library⁵. For each document, we obtained the corresponding source and target versions from publicly accessible materials (licensed under CC-BY-SA 3.0 IGO⁶). These documents are already aligned at document level using the metadata. We sampled five source documents and extracted ten pages per document, each beginning at a chapter header, resulting in a total of 50 pages in English. On this corpus, we created human gold annotations at both the chunk level and sentence level. Two annotators marked boundaries at the chapter or section granularity, and aligned corresponding source-target sentences. Source and target sentences can exhibit many-to-many correspondences. These annotations are used to evaluate chunk detection, chunk alignment, and downstream sentence alignment. Devel-

opment dataset was also constructed following the same manner. (Appendix A)

To assess robustness to non-linear alignments, such as cases where the document structure in the target language differs from that of the source, we also conducted experiments on a shuffled dataset. Such non-linear correspondences can arise when editors deliberately restructure translated editions, especially in abridged translations, by omitting, combining, or reordering sections to fit a shorter format. They may also result from page-layout changes for right-to-left scripts or from explanatory notes added to the translation. This dataset was constructed by randomly sampling chunks from the original target Japanese corpus and permuting their order (Table 1).

All code and data sufficient to reproduce the experiments are released in our repository⁷. Due to licensing constraints on embedded images, however, we hide the pictures, logos, illustrations in the documents for this experiment.

4.2 Metrics

We evaluate our approach along three axes that mirror the pipeline: (i) whether chunks are correctly segmented within each monolingual document; (ii) whether source-language chunks are aligned to the expected target-language chunks; and (iii) whether these intermediate steps improve sentence-level alignment. We evaluate chunk detection from two perspectives: boundary accuracy and content consistency. Following metrics are used.

Boundary Accuracy. We compare predicted chunk boundary positions against human gold boundaries. For each monolingual document, we predict chunk boundaries by identifying the initial sentence of each chunk. A predicted boundary is

²<https://github.com/tesseract-ocr/tesseract>

³<https://github.com/kotaro-kinoshita/yomitoku>

⁴<https://huggingface.co/facebook/nllb-200-distilled-600M>

⁵<https://unesdoc.unesco.org/>

⁶<https://creativecommons.org/licenses/by-sa/3.0/igo/deed.en>

⁷<https://github.com/mkino17/layout-ca>

Model	EN: DocLayNet		JA: DocLayNet+	
	mAP50	50:90	mAP50	50:90
YOLOv11s	0.928	0.755	0.877	0.705
YOLOv12s	0.928	0.765	0.910	0.762
YOLOv12m	0.935	0.773	0.925	0.731
Detailed performance of YOLOv12s				
Title	0.895	0.785	0.995	0.839
Section-header	0.963	0.653	0.908	0.743
Text	0.967	0.838	0.943	0.860
List-item	0.946	0.818	0.955	0.853
Formula	0.836	0.661	0.954	0.797
Picture	0.939	0.867	0.851	0.746
Table	0.932	0.879	0.976	0.925
Caption	0.956	0.878	0.816	0.732
Footnote	0.853	0.647	0.745	0.449
Page-header	0.959	0.699	0.904	0.727
Page-footer	0.966	0.688	0.967	0.710

Table 2: Model comparison – object detection performance across train datasets.

counted as a true positive if it exactly matches a gold boundary position; otherwise it is counted as a false positive, and any gold boundary that is not predicted is counted as a false negative. Based on these counts, we compute precision, recall, and F1.

Average Chunk-level BLEU. Boundary-based metrics alone do not capture whether the content of a predicted chunk matches that of the gold chunk. We additionally evaluate content consistency using Average Chunk-level BLEU, adapted from BLEU-based evaluations (Papineni et al., 2002) used for page-level reading order (Wang et al., 2021). For each gold chunk c in chunks C of the document, we identify the predicted chunk $\hat{c}$ that contains the first sentence (boundary text) of c .

$$\text{Avg. Chunk-level BLEU} = \frac{1}{|C|} \sum_{c \in C} \text{BLEU}(c, \hat{c})$$

For chunk alignment, we anchor evaluation on the gold chunk structure and assess whether the system retrieves the correct target chunk(s) for each gold source chunk. Concretely, for each gold aligned pair $(c^{(s)}, c^{(t)})$, we map the gold source chunk $c^{(s)}$ to its corresponding predicted source chunk $\hat{c}^{(s)}$ with the same rule as chunk detection.

Label	w		
Title	-0.0002	Table	0.3809
Section-header	2.2551	Caption	-0.0002
Text	1.9326	Footnote	0.0657
List-item	0.5171	Page-header	0.5099
Formula	-0.0002	Page-footer	1.1570
Picture	-0.0002	b	-5.4418

Table 3: Weights for chunk similarity score function.

We then take the predicted matches for $\hat{c}^{(s)}$ and obtain the predicted target span $\hat{c}^{(t)}$. If the system produces a one-to-many match, we treat it as a target span composed of multiple consecutive predicted chunks and concatenate their texts. We evaluate whether $\hat{c}^{(t)}$ corresponds to the gold target chunk $c^{(t)}$ using the same two metrics as above: boundary accuracy and average chunk-level BLEU. As for boundary accuracy of chunk alignment, a one-to-many match contributes multiple boundary sentences.

Finally, for sentence alignment, we measure the gain attributable to the chunk layer by comparing a baseline sentence aligner with and without chunk-based preprocessing. We report sentence-level precision, recall, and F1 under both settings and highlight the relative improvement. All metrics are aggregated as document-level means, with sentence matches evaluated under a strict exact-match criterion.

5 Results and Discussions

5.1 Model Parameters

Table 2 and Table 3 present the object detection performance and the learned weights for chunk alignment, respectively. As shown in Table 3, some labels contribute little to alignment performance (largely because they do not appear in the development data), whereas labels such as *Section-header* and *Text* play a central role in aligning source and target chunks.

5.2 Chunk Detection and Chunk Alignment

Figure 3 illustrates how the number of split chunks in English and Japanese varies as a function of the salient ratio r . Although the optimal value of r depends on document type, the UNESCO dataset exhibits a clear tendency around $r \approx 1.3$. At this value, the chunk boundaries produced by our

Evaluation dataset	Number of chunks (pred/gold)	Boundary Accuracy			BLEU
		P	R	F1	
UNESCO (en)	23/18	0.739	0.944	0.829	0.755
UNESCO (ja)	18/18	0.889	0.889	0.889	0.201
UNESCO (shuffled, ja)	18/18	0.889	0.889	0.889	0.201

Table 4: Evaluation results for chunk detection (calculated with $r = 1.3$, chunk similarity score thresh. 0.3).

Evaluation dataset	Chunk pairs (pred/gold)	Boundary Accuracy			Avg. Chunk-level BLEU
		P	R	F1	
UNESCO	15/18	0.737	0.778	0.757	0.184
UNESCO (shuffled)	14/18	0.778	0.778	0.778	0.167

Table 5: Evaluation results for chunk alignment (calculated with $r = 1.3$, chunk similarity score threshold 0.3).

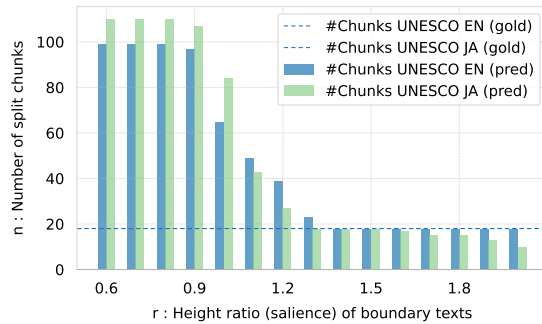


Figure 3: Relationship between salience r and chunk counts. Predicted chunks for English (blue), Japanese (green), and gold counts (horizontal lines).

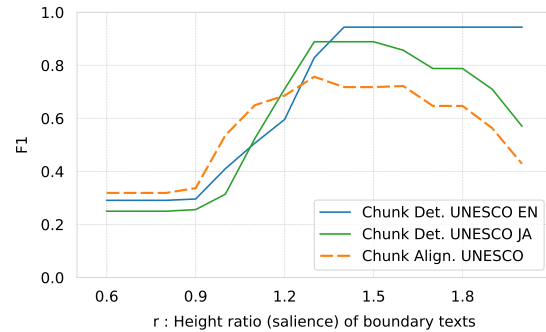


Figure 4: Relationship between salience r and F1 score. Chunk detection English (blue), Japanese (green), and chunk alignment (dashed orange).

method closely approximate the gold-standard segmentation based on chapters and sections.

Figure 4 shows the variation in F1 scores for chunk detection and chunk alignment across different values of r . The horizontal axes of Figures 3 and 4 are scaled identically to facilitate direct comparison. As these figures indicate, under-splitting conditions ($r < 1.0$) result in low precision, while excessive over-splitting ($r > 1.8$) also leads to a consistent decline in F1 scores. These results demonstrate that chunking at an appropriate granularity, corresponding to semantically coherent units, is crucial for achieving robust performance in both chunk detection and alignment.

Tables 4 and 5 present the results of chunk detection and chunk alignment, respectively, at $r = 1.3$, comparing performance across different data-

sets. Comparable chunk alignment scores and average chunk-level BLEU are observed between shuffled and non-shuffled documents, indicating that the proposed method maintains robust alignment performance even when the source and target documents differ substantially in document structure.

An important observation is that a simple layout-based salience measure can approximate discourse-level structure to a meaningful extent. In the UNESCO dataset, setting the salient ratio to $r \approx 1.3$ yields chunks that closely align with chapter- or section-based gold standards, with headers whose height is approximately 1.3 times that of the average body text treated as chunk boundaries. Under-splitting ($r > 1.8$) merges semantically heterogeneous content into a single unit, resulting in am-

Model	UNESCO				UNESCO (shuffled)			
	Sent. pairs (pred/gold)	P	R	F1	Sent. pairs	P	R	F1
VecAlign	976/1409	0.049	0.034	0.040	867/1409	0.030	0.019	0.023
VecAlign (w/ Layout-CA)	1197/1409	0.311	0.263	0.285	1197/1409	0.311	0.263	0.285
BleuAlign	913/1409	0.057	0.037	0.045	629/1409	0.079	0.036	0.049
BleuAlign (w/ Layout-CA)	1155/1409	0.355	0.290	0.319	1155/1409	0.355	0.290	0.319

Table 6: Sentence alignment performance with and without chunk alignment (calculated with $r = 1.3$, chunk similarity score threshold 0.3).

biguous alignment candidates and low precision. In contrast, excessive over-splitting ($r < 1.0$) fragments coherent discourse units and weakens contextual signals, which also leads to a degradation in F1 scores. Although this approach does not explicitly model discourse structure, it suggests that visual prominence in document layout can serve as a weak but effective proxy for semantic organization, particularly in heterogeneous or noisy document settings.

At the same time, the optimal value of r should not be considered universal. The sensitivity of both chunk detection and alignment performance to over- and under-splitting implies that the appropriate granularity is likely to vary across document types, layout conventions, and scripts. Treating the salient ratio as a tunable structural parameter, rather than a fixed heuristic, therefore offers a flexible design choice that can adapt to different datasets.

5.3 Improvement of Sentence Alignment

Table 6 shows sentence alignment performance with and without chunk alignment calculated with $r = 1.3$. The absolute scores are relatively small, reflecting the use of strict matching criteria and the effects of upstream processes, but sentence alignment performance consistently improves when Layout-CA is applied prior to alignment for both Bleualign and Vecalign. This indicates that structurally informed chunk alignment provides a more reliable context for downstream sentence-level alignment.

Without Layout-CA, shuffling of target document substantially reduces the number of aligned sentence pairs, indicating a breakdown of correspondence under disrupted document order. In contrast, when Layout-CA is applied, a similar number of sentence alignments is recovered in both shuffled and non-shuffled settings. This effect is ob-

served for both Vecalign and Bleualign, which implicitly assume a linear, order-preserving correspondence between source and target texts and are therefore more susceptible to ordering mismatches. In such settings, directly aligning sentences without prior structural normalization tends to be less stable, whereas chunk-level alignment effectively constrains the search space and restores meaningful correspondence between semantically related chunks.

6 Conclusion

This paper proposed a layout-aware approach to document-level alignment that integrates chunk detection, chunk alignment, and sentence alignment. Overall, the results indicate that the proposed framework is effective across the alignment pipeline, from extracting coherent chunks to matching them across languages and enabling reliable sentence-level alignment. The results demonstrate that selecting an appropriate chunk granularity is essential for stable alignment performance, as both under- and over-splitting consistently degrade quality. Robust results across shuffled and non-shuffled settings indicate that the proposed method captures structural correspondences beyond surface-level ordering similarity. Sentence alignment further benefits from this structural normalization, with improvements observed for both Bleualign and Vecalign under shuffled conditions, highlighting the effectiveness of the approach in document image settings where ordering noise is common.

Limitations

In this study, we present an initial attempt to incorporate layout-aware chunk alignment into bilingual document alignment from document images, and the

evaluation is therefore limited in scope. Specifically, experiments are conducted on a restricted set of genres and a single language pair (English UNESCO reports and their Japanese translations), and the extent to which the approach generalizes to other document types or low-resource settings remains to be validated. Extending the range of languages and document genres is an important direction for future work.

The main purpose of this study was to test whether an intermediate chunk-level alignment stage improves sentence alignment. We therefore used a modular pipeline and did not optimize the full system for efficiency, measure its runtime or resource cost, or compare it with end-to-end vision-language models. Future work should examine whether the same structural insight can be incorporated into more efficient end-to-end or prompt-based multimodal methods.

Acknowledgements

This work was supported by JST SPRING, Japan Grant Number JPMJSP2180.

References

- Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In Baker, M., Francis, G., and Tognini-Bonelli, E., editors, *Text and Technology: In Honour of John Sinclair*, pages 233–250. John Benjamins, Amsterdam.
- Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In Somers, H., editor, *Terminology, LSP and translation: Studies in language engineering in honour of Juan C. Sager*, pages 175–186. John Benjamins, Amsterdam.
- Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarriás, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., and Zaragoza, J. (2020). ParaCrawl: Web-scale acquisition of parallel corpora. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.
- Baorong, H. (2021). Translation for professionals: corpus-based study of translation universals in computing. In Hu, K., Kim, J.-B., Zong, C., and Chersoni, E., editors, *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 80–88, Shanghai, China. Association for Computational Linguistics.
- Duarte, A. V., Marques, J. D., Graça, M., Freire, M., Li, L., and Oliveira, A. L. (2024). LumberChunker: Long-form narrative document segmentation. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6473–6486, Miami, Florida, USA. Association for Computational Linguistics.
- Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.
- George, N. and Seth, S. (1984). Hierarchical image representation with application to optically scanned documents. In *Proceedings of 7th International Conference on Pattern Recognition (ICPR)*, pages 347–349.
- Ha, J., Haralick, R., and Phillips, I. (1995). Recursive x-y cut using bounding boxes of connected components. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pages 952–955 vol.2.
- Hao, L., Gao, L., Yi, X., and Tang, Z. (2016). A table detection method for pdf documents based on convolutional neural networks. In *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, pages 287–292.
- Koehn, P. (2024). Neural methods for aligning large-scale parallel corpora from the web for south and East Asian languages. In Haddow, B., Koehn, P., and Monz, C., editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1454–1466, Miami, Florida, USA. Association for Computational Linguistics.
- Liu, L. and Zhu, M. (2022). Bertalign: Improved word embedding-based sentence alignment for chinese–english parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.
- Liu, S., Li, Y., and Wei, J. (2025). Xy-cut++: Advanced layout ordering via hierarchical mask mechanism on a novel benchmark.

- Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In Richardson, S. D., editor, *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 135–144, Tiburon, USA. Springer.
- Morishita, M., Suzuki, J., and Nagata, M. (2020). JParaCrawl: A large scale web-based English-Japanese parallel corpus. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France. European Language Resources Association.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Pfitzmann, B., Auer, C., Dolfi, M., Nassar, A. S., and Staar, P. (2022). Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3743–3751, New York, NY, USA. Association for Computing Machinery.
- Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *American Journal of Computational Linguistics*, 29(3):349–380.
- Schreiber, S., Agne, S., Wolf, I., Dengel, A., and Ahmed, S. (2017). Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 1162–1167.
- Sennrich, R. and Volk, M. (2011). Iterative, mt-based sentence alignment of parallel texts. In *Nordic Conference of Computational Linguistics*.
- Soto, C. and Yoo, S. (2019). Visual detection with context for document layout analysis. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *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 3464–3470, Hong Kong, China. Association for Computational Linguistics.
- Team, N., Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. (2022). No language left behind: Scaling human-centered machine translation.
- Thompson, B. and Koehn, P. (2019). Vecalign: Improved sentence alignment in linear time and space. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.
- Tian, Y., Ye, Q., and Doermann, D. (2025). Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*.
- Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2007). *Parallel corpora for medium density languages*, pages 247–258. John Benjamins Publishing Company.
- Wang, Z., Xu, Y., Cui, L., Shang, J., and Wei, F. (2021). LayoutReader: Pre-training of text and layout for reading order detection. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *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.
- Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., and Zhou, M. (2020). Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference*

ence on Knowledge Discovery & Data Mining, KDD '20, page 1192–1200, New York, NY, USA. Association for Computing Machinery.

Yang, X., Yumer, E., Asente, P., Kralej, M., Kifer, D., and Giles, C. L. (2017). Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4342–4351.

Zhao, J., Ji, Z., Feng, Y., Qi, P., Niu, S., Tang, B., Xiong, F., and Li, Z. (2025). Meta-chunking: Learning text segmentation and semantic completion via logical perception.

Zhong, X., Tang, J., and Yepes, A. J. (2019). Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE.

A Dataset

Development

- UNESCO. 2018. Journalism, ‘Fake News’ & Disinformation Handbook for Journalism Education and Training. Paris, UNESCO.

- UNESCO. 2020. Education for Sustainable Development – A roadmap. Paris, UNESCO.

Evaluation

- UNESCO. 2018. Global Education Monitoring Report Summary 2019: Migration, Displacement and Education – Building Bridges, not Walls. Paris, UNESCO.
- UNESCO. 2020. Global Education Monitoring Report Summary 2020: Inclusion and education: All means all. Paris, UNESCO.
- UNESCO. 2021. Global Education Monitoring Report 2021/2: Non-state actors in education: Who chooses? Who loses? Paris, UNESCO.
- UNESCO. 2023. Global Education Monitoring Report Summary 2023: Technology in education: A tool on whose terms? Paris, UNESCO.
- UNESCO. 2024. Global Education Monitoring Report Summary 2024/5 – Leadership in education: Lead for learning. Paris, UNESCO.

Crosslingual Disparities in LLM Performance: Challenges for MT as Mitigation

Rebecca Knowles
Cyril Goutte

Rebecca.Knowles@nrc-cnrc.gc.ca
Cyril.Goutte@nrc-cnrc.gc.ca

Digital Technologies Research Centre, National Research Council of Canada, Ottawa, ON, Canada

Abstract

We examine crosslingual performance disparities in large language models (LLMs) in the context of safety- and regulation-related queries in Canada. We manually build a set of English and French query pairs with gold standard answers and collect LLM-generated answers, which are manually annotated for correctness. We find that LLMs are more likely to produce errors in their answers in French than in English. We investigate a machine translation pipeline, translating the French query, producing an English LLM response, and translating the response back to French. We find that, while it can mitigate some of these performance disparities, additional challenges such as the reliability and language of the cited sources or technical terms greatly impact that mitigation strategy.

1 Introduction

Large language models (LLMs) are known to exhibit crosslingual disparities (Lai et al., 2023; Shen et al., 2024; Wang et al., 2024; Shafayat et al., 2024; Chataigner et al., 2024; Qiu et al., 2023, i.a.): worse performance in languages other than the predominant language on which the model was trained (typically English). When deploying these models in multilingual contexts, this question of crosslingual disparities comes to the forefront. We situate our work in the Canadian context of official bilingualism, focusing on queries about topics related to safety and regulations that users in Canada might seek answers to. Even in this relatively high-resource context—with both English and French having large amounts of monolingual data available and Canadian French–English parallel data also being widely available through Canadian publications online—we show that LLMs are statistically significantly more likely to produce errors in French than in English. We examine a common mitigation for

this: a machine translation (MT) pipeline that translates the question to English, and translates the LLM response back to French. Though the MT pipeline does decrease errors compared to French answers, it introduces other challenges that we examine, particularly around technical content and sourcing. As LLMs improve, in both their overall accuracy and in their ability to perform crosslingually, we expect that the magnitude of these disparities will continue to decrease. Nevertheless, the continued presence of crosslingual disparities, even for high resource languages, also hints at the greater disparities that can be expected (and which are likely harder to mitigate with MT) for lower-resource languages.¹

We describe how we build a small, targeted dataset of English–French query pairs for identifying and evaluating crosslingual performance disparities, including answers produced by three LLMs, manually annotated for correctness (Section 3). The resulting performance disparities are analyzed in Section 4, comparing French and English answers, as well as the MT mitigation strategy. We also

¹See, i.a., Nejadgholi et al. (2024), who used MT tools to translate the query “As a newcomer to Canada, what vaccines should my children get?” into 32 languages commonly used by newcomers to Canada, submitted those translated queries to ChatGPT, and found that prompts in some languages saw 0% LLM accuracy in responding to the query.

discuss the challenges that arise in such an approach. This paper builds on work described in Knowles et al. (2026), which examined a smaller set of queries and three small LLMs.

We make the data available at <https://github.com/nrc-cnrc/DispXLing>.

2 Related Work and Motivations

We begin by situating our work in the context of related work on crosslingual evaluation of LLMs.

A wide variety of evaluation sets for LLMs exists; see Liu et al. (2024) for a more extensive survey, including multilingual evaluation datasets. Within that space, some are designed for multiple natural language processing tasks (Ahuja et al., 2023, i.a.), while others focus on specialized topics like exam questions or math (Zhang et al., 2023; Lai and Nissim, 2024, i.a.). Some are used to evaluate research questions pertaining specifically to multilinguality, such as whether models respond in the same language as the query (Holtermann et al., 2024), or differences in accuracy or truthfulness across languages (Liu et al., 2025). Our query set is relatively small, and designed for the particular context of Canada and its official languages, French and English.

While it is increasingly common to scale up evaluation through the use of LLM-as-a-judge evaluations, we rely exclusively on *human* evaluation. This is motivated by our focus on crosslingual performance disparities. Since we want to know whether the LLMs we evaluate exhibit crosslingual performance disparities, it is not safe to also use them for evaluation. An LLM-as-a-judge evaluation that showed either worse or better performance on French could be attributable to the *generating* model’s (in)ability to answer the query *or* to the *evaluating* model’s (in)ability to accurately judge the output, without a clear way to untangle these. We acknowledge that performing human evaluation limits the scale of our evaluation, but it eliminates the confounding factor of potential crosslingual disparities in LLM-based evaluation.

One common mitigation for LLM crosslingual disparities is the use of machine translation (Section 3.4). MEGA (Ahuja et al., 2023) evaluates a translate-test approach, where the source language question is translated to English, submitted to the LLM, and the resulting response eval-

uated. They find improvements for low-resource languages, but not for French. M3Exam (Zhang et al., 2023) also finds mixed results in a translate-test paradigm. Other work explored approaches for crosslingual “thought prompting” (Huang et al., 2023), by prompting the model to first translate the input into English before generating the response. A related approach (Wang et al., 2024) uses “Please think in English and then generate the response in the original language.” as instruction to the model. We use an approach close to the translate-test approach, though we complete the pipeline by translating the resulting English response back to French for evaluation; we refer to this as FR-EN-FR. Our approach should be viewed as a simulation of how one might deploy translate-test in a scenario with real users querying in French and expecting answers in French.

3 Data and Methods

The following subsections describe our approach in detail. In Section 3.1, we describe the design and creation of the query set and corresponding answers that form the backbone of our analysis. The three LLMs we choose to evaluate are described in Section 3.2. Section 3.3 describes our manual annotation protocol. Finally, the MT mitigation approach is detailed in Section 3.4.

3.1 Query Set

Our query set consists of 132 manually-produced English–French query pairs (with corresponding gold standard responses), based on Canadian provincial and territorial government webpages and other authoritative documents (see Appendix A for additional details). The queries were initially written in English and translated by fluent French speakers into French. This was followed by additional checking of the query pairs by fluent speakers of French and English with backgrounds in translation, and queries in both languages were adjusted to improve consistency.

Of the 132 pairs, 82 are focused on vaccination in Ontario, with the remaining queries split evenly (10 query pairs apiece) across five other topics: winter tires (Quebec), driving licences (British Columbia), health cards (Alberta), driving and rules of the road (New Brunswick), and hunting and fishing (Nunavut). We chose query topics that cov-

ered a variety of geographical areas of Canada, a linguistically-diverse set of regions, as well as a variety of safety and regulatory topics that might be of interest to LLM users. In most cases, bilingual English and French information was available online.² As the sources are available on the internet, they may be found in the LLMs’ training data.³

There are a number of ways we could have chosen queries, such as by selecting them from online FAQs (at the risk of training data contamination if the full FAQs appeared in the model training data), by collecting real queries from potential users, and so on. In manually creating this set of queries, we tried to balance plausible topics of queries with the constraint of being able to answer them directly and concisely with minimal ambiguity using the relevant authoritative source. As a consequence, some of the queries are likely to be substantially more specific than the average user might produce; i.e., a user might be more likely to ask “What vaccines should babies get?” (not in our query set) than *According to the routine immunization schedule in Ontario (Canada), what vaccines should babies receive at 15 months old?* (in our query set). The latter question has a more clearly defined answer (making it easier to annotate), while the correctness of the answer to the former strongly depends on the context in which it is asked and how it is interpreted.

<i>Query:</i>	When should children get the DTaP-IPV-Hib vaccine in Ontario (Canada)?
<i>Gold:</i>	2 months
	4 months
	6 months
	18 months

Table 1: Example query and gold sub-answers.

We manually curate a set of gold standard sub-answers for each query, breaking up the response into something comparable to the “atomic facts” described in Min et al. (2023). For example, as shown in Table 1 for the English query *When should children get the DTaP-IPV-Hib vaccine in Ontario (Canada)?*, the correct response consists of four

parts (sub-answers): *2 months, 4 months, 6 months and 18 months*. The corresponding French query, *Quand les enfants devraient-ils recevoir le vaccin DCT-polio-Hib en Ontario (Canada)?*, has its response broken down into the same corresponding responses: *2 mois, 4 mois, 6 mois, and 18 mois*. The queries were designed with the intention of having responses that could be easily broken up into these sub-answers. Our query set is a refined version of a query set that we initially tested with smaller LLMs (Knowles et al., 2026); we used that initial set of experiments to refine both the query set (including an attempt to remove as much ambiguity as possible) and to refine our annotation protocol. An example of this is the addition of the clarifying “(Canada)” in our queries, as we originally found some responses that might have been explainable if, for example, Ontario was erroneously interpreted as Ontario, California. Nevertheless, in some cases, we found through the annotation process that there were still ambiguities or underspecifications. The ground truth answers were designed for human annotation rather than automatic evaluation; in some cases, we provided additional parenthetical information for the human annotators in the gold sub-answers.

3.2 Large Language Models

We analyze the responses of three LLMs: GPT 4.1 (model version: 2025-04-14), GPT 5 Chat (model version: 2025-08-07), and Aya 23 (35B).⁴ We are limited in the number of models we analyze by the constraints of manual annotation; the specific results we observe in this work should not be assumed to generalize to all models or all language pairs, but the risk that they may generalize motivates this line of research and supports the argument for continued crosslingual evaluation before model and tool deployment.

Two are widely used by the public and for building LLM-based tools; the third is smaller, explicitly-multilingual, and deployed locally. The GPT models are accessed via Azure’s OpenAI API (Microsoft Foundry). In order to obtain similar behaviour to ChatGPT, we use chat completions and

²The British Columbia Graduated Licensing Program information is available in English, Simplified Chinese, and Punjabi, but not French. For that category, we rely on human translation without the authoritative French source; the situation is similar for the hunting source for Nunavut, but we were able to use the French version of the fishing source to verify some details.

³Some are more recent than the training date, but they are often unchanged (or have only limited changes) from previous versions.

⁴Model fingerprints, respectively: fp_7a7fd0eb44, fp_88bf7c189b and fp_ad66631dec, bab44e009440

set the system message to “Assistant is a large language model trained by OpenAI.”⁵ We set seed=1 for improved replicability (deterministic output is not guaranteed), temperature to 0.0, and maximum number of tokens to 1000.

The Aya 23 (35B) model (henceforth Aya:35B) is an explicitly multilingual model produced by Cohere (Aryabumi et al., 2024). We use Ollama⁶ for decoding, with the generate interface, seed=1, temperature of 0.0, and 1000 tokens maximum. We use the default quantized version (Q4_0) to run locally and efficiently. The substantial difference in size should be taken into account when examining results of this model and the GPT models, as the smaller model’s results in terms of accuracy are expected to be lower. There are several reasons to expect lower accuracy: the age of the model and its training data (though the answers to many of the queries posed in our experiments have not changed in this time span), the use of quantization, and the smaller size of the model.

3.3 Annotation Protocol

We use the brat⁷ annotation tool (Stenetorp et al., 2012) to perform annotations on each of the query-response pairs (one query, one LLM response). In the interface, annotators are presented with the query, its gold standard sub-answer(s), and an LLM response. As the main focus of this work is on the factual accuracy of the responses to the queries, we ask the annotators to apply the following labels: CORRECT (a response that is semantically equivalent to one of the gold standard responses), PARTIAL (one that is partially correct, e.g., matches one half but not the other half of a span of time), and WRONG (one that does not match the gold responses or that contradicts the gold responses). The CORRECT and PARTIAL labels are then manually linked to the GOLD sub-answers as part of the annotation process; it is these linked pairs (and the unlinked WRONG answers) that we analyze in the remainder of the work. Annotators also provided other labels for refusals to answer, irrelevant responses, a lack of direct answer,

or other (along with free text comments on the labels). Prior work (Knowles et al., 2026) found low levels of agreement on PARTIAL labels, discarding them rather than examining them. In our initial attempt to improve agreement and thus enable us to use the information from PARTIAL labels, we provided more extensive instructions and examples of when to use the PARTIAL label, including specifying that it should be used when a vaccine name is close but not exact (such as the distinction between Tdap and DTaP); in practice we observed that this was still challenging in the vaccine setting, with some annotators not using the PARTIAL label in the expected locations, so we added an additional step for those query-response pairs (see Section 4.3).

3.4 Machine Translation

For our experiments in mitigating crosslingual disparities, we first used machine translation to translate the French queries to English. All 132 queries differed at least slightly between the original English and the machine-translated English. We then submitted these machine-translated queries to each of the LLMs, producing LLM responses, which we in turn machine translated into French. All machine translation was performed using the translation function (a call to Microsoft Translator models)⁸ in Microsoft 365’s Excel between March and April 2026, using the language options of “en” (English) and “fr-ca” (Canadian French).⁹ In light of our use of Microsoft Foundry to query both GPT models, it is a realistic choice to use a Microsoft-provided translation solution, as someone implementing this procedure realistically might also choose to use translation tools accessible through the same providers. We perform a manual comparison of the original English queries and the queries automatically translated from French into English (MT-EN). In three cases, the MT-EN queries use more formal terms for diseases than the original English (herpes zoster for shingles, and pertussis for whooping cough, twice), and on two occasions there is a mistranslation of a vaccine name (“DTaP-

⁵<https://learn.microsoft.com/en-us/azure/ai-foundry/openai/how-to/chatgpt?view=foundation-classic&source=recommendations&tabs=python-secure#basic-example>

⁶<https://ollama.com/>; version 0.15.0

⁷Specifically, we use commit 44ecd825810167eed2a5d8ad875d832218e734e8, the most recent version from <https://github.com/nlplab/brat>, which permits improved visualizations of arrow wrapping across multiple lines.

⁸<https://support.microsoft.com/en-us/excel/translate-function>

⁹In one case, an LLM response was too long to translate and had to be split in two; all others were translated as a single text.

IPV-Hib” mistranslated as “DPT-polio-Hib” instead of “DCT-polio-Hib”).¹⁰ Among the most frequent differences in the MT-EN queries are “Ontario (Canada)” being reformatted as “Ontario, Canada”, rephrasings (“months old” vs. “months of age” or “teens” vs. “adolescents”), reframing of queries that should not change the answer (“Combien de jours durant l’année faut-il [...]” being translated as “How many days in the year do I need to [...]” rather than “How many days in a year do you have to [...])”. We observe few grammatical errors (one case of subject-verb disagreement) and some occasional awkward phrasings (dropping the article “the” before vaccine names).

4 Results

We now describe the results of our study, beginning with basic text statistics (Section 4.1) and annotator statistics (Section 4.2). We then describe how we derived query-level labels, by combining annotations from several annotators through majority voting (Section 4.3). Our results allow us to compare English and French results (Section 4.4) and explore how the MT pipeline mitigates performance disparities (Section 4.5), before examining URLs and other assorted challenges (Sections 4.6 and 4.7).

4.1 Text Statistics

The average query length for English is 13.9 tokens (minimum of 8, maximum of 26).¹¹ For French, the average query length is 15 tokens (minimum of 9, maximum of 33). The queries differ in terms of their numbers of gold standard sub-answers as well. Every query has at least one sub-answer, with some having as many as 6 sub-answers (mean of 2 across all queries). The vaccine-related queries have a larger average number of sub-answers (2.6) as compared to the non-vaccine queries (1.1), with the latter having a maximum number of 3 sub-answers.

We also examine the response lengths by model and language of query and response (see Table 2). Both of the GPT models tend to produce long answers, often with multiple sections and with sum-

maries and supposed sources (see Section 4.6 for discussion of the validity of the linked sources). The Aya:35B model sometimes produces extremely short answers (one or two words in response to yes or no questions, for example).

Across all of our experiments, the LLM responses were in the expected language (i.e., in English for English-language queries and in French for French-language queries). Aside from a few minor MT errors in the translation pipeline output, our annotators did not report specific linguistic problems with the French or English output.

4.2 Annotator Statistics

Three annotators performed English language annotations and six annotators performed French language annotations (including one who contributed to both languages). English language annotators annotated between 246 and 396 query-response pairs (the full set), for an average of 298 query-response pair annotations. For French, with a larger number of annotators, each annotator annotated between 138 and 261 query-response pairs, for an average of 225.5 query-response pair annotations. For FR-EN-FR, three of the French language annotators participated, with their numbers of annotations ranging from 87 to 276 with an average of 205.5.

All annotators completed annotations (French or English) of one particular subset (“vaccine-ontario-realistic”) of the vaccine queries; this gives us a controlled set of 11 query pairs on which to compute interannotator agreement. We compute pairwise interannotator agreement (IAA; Table 3) using the scikit-learn (Pedregosa et al., 2011) implementation of Cohen’s κ (Cohen, 1960).¹²

Label	English			French		
	Min.	Mean	Max.	Min.	Mean	Max.
CORRECT	0.707	0.792	0.911	0.429	0.673	0.861
PARTIAL	0.298	0.467	0.639	0.081	0.303	0.728
WRONG	0.814	0.875	0.936	0.579	0.760	0.933

Table 3: Pairwise Cohen’s κ on vaccine-ontario-realistic data subset.

¹⁰Other MT systems (Google Translate and DeepL) make similar errors in brief experiments with several different contexts: leaving “DTaP-IPV-Hib” unchanged (English to French), or changing “DCT-polio-Hib” to “DTaP-polio-Hib” (French to English). We note that Ontario’s Routine Immunization Schedule differs from the National Vaccine Catalogue (<https://nvc-cnv.canada.ca/en/vaccine-catalogue>) in the former’s use of “DCT-polio-Hib” rather than “DCaT-VPI-Hib” in French.

¹¹For simplicity, we use whitespace-separated token counts, not subword segmentations.

¹²https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

Model	English			French			FR-EN-FR		
	Min.	Mean	Max.	Min.	Mean	Max.	Min.	Mean	Max.
GPT 4.1	48	158.5	390	24	168.6	338	50	195.5	464
GPT 5 Chat	32	178.0	491	46	202.4	462	42	211.9	512
Aya:35B	2	131.6	325	1	144.8	383	1	128.3	378

Table 2: Response lengths (whitespace-separated tokens) by model and language.

We compute agreement on the CORRECT labels over the set of all 39 sub-answers. For English, the mean pairwise IAA on this query subset is 0.792 and for French it is 0.673, which still constitutes moderate to strong agreement. Interpretations of Cohen’s κ are not widely agreed upon, but under the framework described by Landis and Koch (1977) (which we use for the remainder of this section), these would constitute substantial and moderate agreement, respectively. The vaccine queries were chosen because they tend to be challenging to annotate; the minimum pairwise agreement on the broader set of annotations is larger.

We also compute agreement on the PARTIAL labels; this drops on our controlled set for both English (mean agreement moderate) and French (mean agreement fair). Across the broader dataset this rises somewhat for English (minimum of 0.523) but remains lower for French. We manually investigated this and observed that a substantial portion of this was due to differences in how closely annotators followed the instruction to label incorrect (but similar) vaccine names as PARTIAL; we discuss how we mitigate this mismatch in Section 4.3.

Finally, we compute agreement at the query level of whether or not a WRONG label is applied anywhere in that response. English agreement on WRONG labels is almost perfect and French is moderate to almost perfect. The minimum agreements increase when computed on the broader set of queries.

4.3 Generating Labels

While we observed generally strong agreement between annotators on the CORRECT and WRONG labels, there is still some variation, which may be due to differences in perspective and interpretation or due to human error. For this reason, we collected multiple annotations and assign query-level labels to the query-response pairs based on majority vote. For each of the query’s sub-answers, we observed whether more than half of the annotators agreed that

a CORRECT label should be assigned; we then provided a COMPLETE label only if this was the case for all sub-answers (i.e., the complete correct answer was provided). If this is not the case, a query-level label of INCOMPLETE is assigned.

We also give a query-level label of ERROR or NO-ERROR. The label of ERROR is assigned if more than half of the annotators labeled some part of the LLM answer as WRONG, or if more than half of the annotators gave a PARTIAL label to one (or more) of the sub-answers. Since we noted the inconsistencies in vaccine name annotation, we add a third way in which an ERROR label can be assigned: for the two vaccine query subsets that ask for vaccine names in the response, we performed a separate annotation in which we collected a set of “confusables” (similar-sounding vaccine names that refer to different vaccines, such as Td/Tdap/DTaP/DTap-IPV/DTaP-IPV-Hib, MMR/MMRV, Men-C-C/Men-C-ACYW), automatically labelled those in the responses, and manually checked all of these to see whether they were indeed directly answering the query *and* ought to have been labelled as PARTIAL. If one appeared, the label of WRONG was also applied. This mitigates one of the major sources of inconsistency between annotators (particularly in French, but also in English).

Each query-response pair receives a combined label: COMPLETE/NO-ERROR, COMPLETE/ERROR, INCOMPLETE/NO-ERROR, or INCOMPLETE/ERROR. Their interpretations are shown in Table 4. The ideal result is COMPLETE/NO-ERROR: a response that correctly answers the query with no wrong information. INCOMPLETE/ERROR is arguably the worst result: it does not completely answer the query and it gives at least one wrong answer. The other two are more complex but generally fall in between the extremes: a COMPLETE/ERROR response does include all of the components of a correct answer, but also provides incorrect information, while a

INCOMPLETE/NO-EERROR response doesn't provide the full correct answer, but it also doesn't provide wrong information (e.g., when the LLM response is a refusal to answer). We did not observe any refusals to answer the queries in this set of experiments. Annotators also provided annotations for "noncommittal" responses from the LLMs; these were quite rare in the data as well, but may not have been as consistently annotated as other labels.

	NO-EERROR	ERROR
COMPLETE	Response contains the full set of correct sub-answers <i>and</i> does not contain wrong answers.	Response contains the full set of correct sub-answers <i>but</i> also contains wrong answers.
INCOMPLETE	Does not contain the full set of correct sub-answers <i>but</i> does not contain wrong answers.	Does not contain the full set of correct sub-answers <i>and</i> also contains wrong answers.

Table 4: Interpretations of query-level label combinations.

4.4 Comparing French and English Results

In Figure 1, we first focus on the English and French language queries and responses and observe that in all cases, the French language responses have more INCOMPLETE/ERROR labels, as well as more of any type of label that contains ERROR than the English queries do. Both GPT 4.1 and GPT 5 Chat have more COMPLETE/NO-EERROR labels in English than in French, and all three models have more COMPLETE labels overall in English. While we show the results for Aya:35B in the same figure for convenience, we note that it is a substantially smaller (and less recent) model than the two GPT models, and as such should not be expected to have the same performance.

Due to the large number of ties, we compute significance using the trinomial test (Bian et al., 2011) as implemented in the python package TrinomialTest (Rudar and Lung, 2025).¹³ We compare the paired English and French queries, and find using a one-sided trinomial test that both GPT 4.1 and GPT 5 Chat are significantly more likely to produce responses that contain an ERROR in French than in English ($p = 0.01$ and $p = 0.006$, respectively),

¹³Version 1.0.4.

rejecting the null hypothesis that there is no difference across languages. For Aya:35B, we cannot reject the null hypothesis ($p = 0.2$), but the overall high rates of ERROR do highlight the challenges with a smaller quantized model. In terms of COMPLETE responses, only GPT 4.1 has statistically significantly more COMPLETE responses in English than in French ($p = 0.03$).

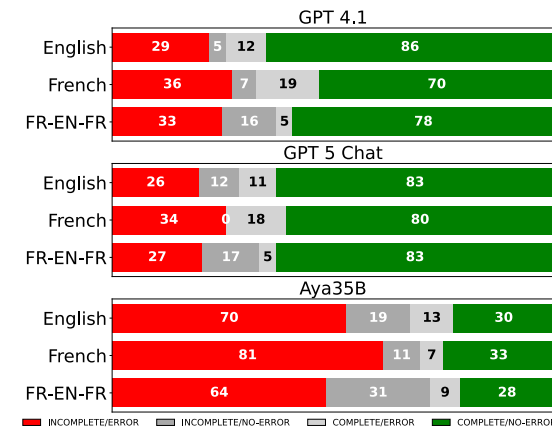


Figure 1: Results across all queries, separated by LLM and language or translation pipeline. From left to right, the bars show INCOMPLETE/ERROR, INCOMPLETE/NO-EERROR, COMPLETE/ERROR, and COMPLETE/NO-EERROR.

We also looked at the results by category of query; the vaccine queries tend to be more difficult for LLMs to produce COMPLETE answers for. This is due in part to the fact that, on average, they have a larger number of sub-answers. When we separate vaccine queries from non-vaccine queries (pooling models; see Figure 2), we see that the non-vaccine queries have equal numbers of COMPLETE responses in French and English, but more ERROR responses in French than English. The vaccine queries have more COMPLETE responses in English than in French and more ERROR responses in French than in English. When we separate out the models while also separating the queries into vaccine and non-vaccine groups, it is always the case that the French has more INCOMPLETE/ERROR responses than the English (though in some cases it has fewer total INCOMPLETE responses). In all but one case, the English has more COMPLETE responses. We do not compute signifi-

cance give the smaller data size.

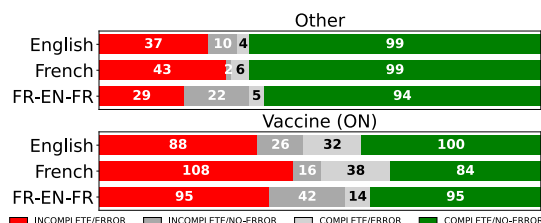


Figure 2: Results across LLMs and language or translation pipeline, divided into vaccine and other (non-vaccine) queries.

4.5 Machine Translation as Mitigation

We see in Figure 1 that the simulated translation pipeline approach does mitigate the performance disparities, with the number of ERROR responses dropping from the French to close to (or in the case of Aya: 35B, below) the level of ERROR labels in English. In no case is there a statistically significantly greater number of ERROR labels in FR-EN-FR than in English. Similarly, the number of COMPLETE labels increases. For GPT 4.1, there are still significantly more CORRECT labels in English than in FR-EN-FR, but this is the only model for which this is the case. At first glance, it appears that machine translation as a mitigation is a clear success, as it does eliminate the gap we observed.

However, if we take a closer look at the data, concerning patterns emerge. Looking at the vaccine name confusables we identified in two of the vaccine query subsets, we observe that out of 66 query-response pairs (22 queries and 3 LLMs), 20 in English had errors due to confusables, as compared to 42 in French and 48 in FR-EN-FR. For all LLMs except Aya: 35B, there are significantly more errors due to confusables (ignoring all other sources of errors) in French than in English ($p = 0.022$ for GPT 4.1, $p = 0.003$ for GPT 5 Chat, and $p = 0.156$ for Aya: 35B) and significantly more errors due to confusables in FR-EN-FR than in English for all LLMs ($p < 0.009$ in all cases). In short, there are more vaccine name errors in French and FR-EN-FR than there are in English. There are several possible sources for these errors. In some cases, the Ontario routine immunization schedule (French version) uses a different translation for vaccine names than other parts of Canada and other parts of the

world; the LLMs may be using a term more commonly found in their training data. The additional errors in vaccine names in FR-EN-FR may be due to errors introduced in the translation process, as in some cases they seem to produce made-up vaccine names. The most confusable vaccine names tend to be highly abbreviated (e.g., DTaP-IPV-Hib), and are likely to be infrequent sequences of subword segments, which have been observed to cause translation errors in neural machine translation systems (Kothur et al., 2018).

Another challenge for the machine translation pipeline is the production of reference URLs, as discussed in more detail below.

4.6 Analysis of URLs

The reliability of provided sources has been identified as a weakness of LLMs. In the context of our study, two of the LLMs (GPT 4.1 and GPT 5 Chat) tend to output URLs as claimed sources or references for more information. Over the 1188 LLM answers, a total of 868 URLs are included, 294 of them unique. We extract URLs by regular expression and use a combination of automated and manual checks (including annotator-provided comments in OTHER) to confirm whether they link to real webpages and whether those are indeed relevant. When the URL actually lands on a valid page, the annotator manually checks its language, and whether the page content is relevant to the topic (e.g. about the right vaccine).

Table 5(a) shows that, overall, less than 30% of the 294 distinct URLs (86/294) lead to a valid page (a 4-to-1 English/French ratio). Ratios are more skewed on the occurrences: only 26% of returned URLs work, with an English-to-French ratio of 85%/15%. This is concerning: although the number of URLs provided is similar over English (313), French (262), and FR-EN-FR (292) answers, there are many fewer valid French URLs. By far, the most common outcome is “not found” (404 error), indicating a working server, but missing page. There are also a number of empty pages and non-responsive (or non-existing) servers. Interestingly, some empty pages, in particular links to PDF documents, are re-used repeatedly, and constitute about 21% of all URLs appearing in LLM replies.

Table 5(b) shows the language of the landing pages for the working URLs, for the three types of

(a)	Status	unique		all	
		<i>n</i>	(%)	<i>n</i>	(%)
	valid (EN)	68	23.1	190	21.9
	valid (FR)	18	6.1	34	3.9
	not found	178	60.5	454	52.3
	empty	26	8.8	186	21.4
	<i>nnp</i>	4	1.4	4	0.5
	Total	294	100.	868	100.

(b)	landing page language			Total
	Query	EN	FR	
	English	92	4	217
	French	33	27	203
	FR-EN-FR	65	3	224
	Total	190	34	644

Table 5: (a) Statistics for unique, and all occurring URLs. *nnp*=server not responding (likely non-existent). We divide the valid URLs by landing page language. (b) Language of URL landing page (when available) for three query strategies (English, French, FR-EN-FR). <NA> are invalid URLs (no language).

queries (English, French and FR-EN-FR pipeline). English queries elicit more URLs, a larger share ($96/313 \approx 31\%$) are valid, and they overwhelmingly land on English pages. By contrast, French queries result in fewer URLs, only about 23% are valid, and, perhaps unsurprisingly, a majority of them do *not* land on French pages. Finally, the MT mitigation approach yields interesting results: only about 23% of the produced URLs are valid, but the ones that do overwhelmingly land on English pages. This means that the queries, once translated, do not produce as many URLs as the original English queries (although the difference is unlikely to be significant), and the translation back to French, although it may help with the correctness of the answer, unsurprisingly does not transform the supporting URLs, in the few times where they work, into links leading to valid French information.

A natural question is whether some of the URLs that are no longer valid did actually exist when LLM training material was obtained. It is difficult to systematically examine this question. However a manual inspection of the most frequent URLs, using the Wayback Machine¹⁴, shows that most cannot be found in the Internet Archive. Out of the 43 most frequently observed unique URLs (69% of all URL mentions), only 7 are found in the Internet Archive, and most of the snapshots are either empty (maybe because they are unarchived PDF files) or unavailable (possibly because they used to exist and were moved elsewhere). For example, the most frequent URL (supposedly Ontario immunization schedule information in English), seen 88 times in LLM answers, can be found with the

Wayback Machine between 2013 and 2024, but is now empty. The French version of that URL (same path, with a */fr/* instead of */en/*), is seen 5 times, also found in the Internet Archive, between 2014 and 2024, and also empty now. Similar URLs are provided in other answers, with slight differences in path or file name (e.g. 21 occurrences of *immunization_schedule.pdf* vs. 19 occurrences of *ImmunizationSchedule.pdf*), pointing to non-existing or empty pages. The current accurate URL¹⁵ was available as early as January 2024, before the training data cutoff date for GPT 4.1.

Our manual inspection may be biased in several ways. For example, we can not rule out that URLs existed previously on sites that disallow crawling (such as the Government of Nunavut site), and are therefore not in the Internet Archive. However, these results suggest that the vast majority of URLs are likely made up, combining parts of existing servers, plausible paths and filenames inspired by the context of the interaction.

4.7 Additional Challenges

While we have examined the results quantitatively so far, we also provide a brief discussion of individual anomalies and other qualitative notes. Our query set was designed with the intention of being as unambiguous as possible, but it is difficult to foresee all possible ambiguities. For example, one GPT 5 Chat response to a question about the novice licences in British Columbia provided a response about obtaining a “novice amateur radio licence” rather than a driver’s licence. This ambiguity was not flagged in our initial round of dataset creation,

¹⁴<https://web.archive.org>, last used June 2026. Note that this free service is occasionally unreliable.

¹⁵<https://www.ontario.ca/page/ontarios-routine-immunization-schedule>, as of this writing.

as no LLMs had exhibited that interpretation before, and the French query was unambiguous. While we attempted to make our gold responses as comprehensive as possible, some of the LLMs produced responses, particularly to vaccine queries, with extensive additional information about unusual circumstances in which the vaccination timeline may differ or for circumstances where catch-up vaccinations were required.

Additionally, as we noted in the analysis of URLs, we observe fabrications of text and sources. This is simpler to verify for the URLs themselves, but we also note fabricated text and quotations; in at least one case, an English response about tire requirements in Quebec produces a non-existent (404 error) but plausible-looking URL followed by a purported quote from that non-existent page, which directly contradicts the actual requirements on the real webpage. Other annotators noted similar problems in French. We expect that a more detailed analysis of other quotations would result in more examples of fabrication or false information.

5 Limitations

The small size of our query set, bounded in part by the time required to do manual evaluation, means that we are limited in the conclusions we can draw. Additionally, the queries themselves are more controlled than a real user's queries would likely be. This decision was made in order to facilitate evaluation. We expect that it likely results in an overestimate of model performance, but have not done the additional experiments that would be required in order to verify this.

In writing an academic paper, rather than deploying a real system, we are afforded some freedom in the choice of queries and how we measure performance. For both the deployment of LLM-based tools and any MT mitigation used to minimize crosslingual disparities, we would recommend taking a user-centred perspective. This raises questions of how to balance factual correctness with the user experience, including what weight to give to the respective benefits and pitfalls of MT mitigation. Our current evaluation does not provide us with clear answers to these questions, and we expect they would vary depending on the user population (e.g., whether users would prefer links to French sources or accept links to English sources). We suggest that those in-

terested in deploying models consult with potential users to better understand their needs and preferences.

As we have discussed, this work is limited both in terms of the size and distribution of the query set, as well as the model selection. The results should not be interpreted as broad claims about the performance of all LLMs or about the performance of certain types of LLMs (e.g., open-weight LLMs, in particular since the open-weight LLM we examined in this work is comparatively small and not as new as the closed-weight LLMs examined). Rather, this should serve as a reminder that these types of evaluations are necessary when considering the deployment of LLM-based tools in multilingual settings; Knowles et al. (2026) showed this to be the case for a set of smaller LLMs, and here we show that, despite overall improvements in accuracy, crosslingual disparities persist in these larger LLMs as well.

6 Conclusions and Future Work

In this work, we use a small tailored query set, focused on French and English in Canadian contexts, to show that crosslingual performance disparities persist in LLMs. We explore a standard MT-as-mitigation pipeline, and find that it does mitigate the statistically significant performance disparities, but it also raises new problems and challenges, particularly regarding technical terminology, links, and purported quotations. We encourage organizations and users considering the deployment of LLM-based tools in multilingual settings to carefully validate and test them for crosslingual performance disparities.

Acknowledgements

This project builds on work that was conducted by the NRC on behalf of the Canadian AI Safety Institute (CAISI). We thank our colleagues for helpful discussion, feedback, and for contributing to the annotation process, and the reviewers for their useful comments and questions.

References

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Ahmed, M., Bali, K., and Sitaram, S. (2023). MEGA: Multilingual evaluation of generative AI. In Bouamor, H., Pino, J.,

- and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.
- Aryabumi, V., Dang, J., Talupuru, D., Dash, S., Cairuz, D., Lin, H., Venkitesh, B., Smith, M., Campos, J. A., Tan, Y. C., Marchisio, K., Bartolo, M., Ruder, S., Locatelli, A., Kreutzer, J., Frosst, N., Gomez, A., Blunsom, P., Fadaee, M., Üstün, A., and Hooker, S. (2024). Aya 23: Open weight releases to further multilingual progress.
- Bian, G., McAleer, M., and Wong, W.-K. (2011). A trinomial test for paired data when there are many ties. *Mathematics and Computers in Simulation*, 81(6):1153–1160.
- Chataigner, C., Taïk, A., and Farnadi, G. (2024). Multilingual hallucination gaps in large language models.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Holtermann, C., Röttger, P., Dill, T., and Lauscher, A. (2024). Evaluating the elementary multilingual capabilities of large language models with MultiQ. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4476–4494, Bangkok, Thailand. Association for Computational Linguistics.
- Huang, H., Tang, T., Zhang, D., Zhao, X., Song, T., Xia, Y., and Wei, F. (2023). Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.
- Knowles, R., Azadi, F., Simard, M., Lo, C.-k., Larkin, S., Tessier, M., and Goutte, C. (2026). Évaluer et atténuer les différences de performance des GLM pour les tâches de génération de textes français et anglais. In *Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, volume 1, pages 238–258, Nantes, France. ATALA.
- Kothur, S. S. R., Knowles, R., and Koehn, P. (2018). Document-level adaptation for neural machine translation. In Birch, A., Finch, A., Luong, T., Neubig, G., and Oda, Y., editors, *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 64–73, Melbourne, Australia. Association for Computational Linguistics.
- Lai, H. and Nissim, M. (2024). mCoT: Multilingual instruction tuning for reasoning consistency in language models. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12012–12026, Bangkok, Thailand. Association for Computational Linguistics.
- Lai, V. D., Ngo, N., Pourn Ben Veyseh, A., Man, H., Deroncourt, F., Bui, T., and Nguyen, T. H. (2023). ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.
- Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Liu, W., Wu, N., Ding, W., Liang, S., Gong, M., and Zhang, D. (2025). Selected languages are all you need for cross-lingual truthfulness transfer. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8963–8978, Abu Dhabi, UAE. Association for Computational Linguistics.
- Liu, Y., Cao, J., Liu, C., Ding, K., and Jin, L. (2024). Datasets for large language models: A comprehensive survey. *arXiv preprint arXiv:2402.18041*.
- Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Nejadgholi, I., Molamohammadi, M., and Bakhtawar, S. (2024). Social and ethical risks posed by general-purpose llms for settling newcomers in canada.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Qiu, Y., Ziser, Y., Korhonen, A., Ponti, E., and Cohen, S. (2023). Detecting and mitigating hallucinations in multilingual summarisation. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8914–8932, Singapore. Association for Computational Linguistics.
- Rudar, J. and Lung, O. (2025). Trinomialtest.
- Shafayat, S., Kim, E., Oh, J., and Oh, A. (2024). Multi-FACT: Assessing factuality of multilingual LLMs using FACTscore. In *First Conference on Language Modeling*.
- Shen, L., Tan, W., Chen, S., Chen, Y., Zhang, J., Xu, H., Zheng, B., Koehn, P., and Khashabi, D. (2024). The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680, Bangkok, Thailand. Association for Computational Linguistics.
- Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). brat: a web-based tool for NLP-assisted text annotation. In Segond, F., editor, *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.
- Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J.-t., Jiao, W., and Lyu, M. (2024). All languages matter: On the multilingual safety of LLMs. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5865–5877, Bangkok, Thailand. Association for Computational Linguistics.
- Zhang, W., Aljunied, M., Gao, C., Chia, Y. K., and Bing, L. (2023). M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Queries

In this appendix, we provide more detailed information about the content of our query set.

- **Vaccines (Ontario):** There are 82 queries related to vaccination in Ontario (primarily children’s vaccinations). These are drawn from Ontario’s Routine Immunization Schedule.¹⁶ Of these, 11 are relatively formal questions of the form “According to the routine immunization schedule in Ontario (Canada), what vaccines should babies receive at [certain ages]?” and “D’après le calendrier de vaccination systématique de l’Ontario (Canada), quels vaccins les bébés devraient-ils recevoir à [...]?”, 11 are less formal versions of the same: “What vaccines should babies in Ontario (Canada) get at [...]?” and “Quels vaccins les bébés devraient-ils recevoir à [...] en Ontario (Canada)?”. The answers to these consist of lists of vaccines. The remaining queries ask about the age at which individuals should get certain vaccines or the age at which individuals should be vaccinated against particular diseases, each again split into two forms (“At what age”/“À quel âge” and “When”/“Quand”). The answers to these consist of lists of ages or age ranges.
- **Winter Tires (Quebec):** There are 10 queries about Quebec’s winter tire regulations,¹⁷ covering topics such as when the regulations apply, to whom and which vehicles they apply, and details of the requirements.
- **Driving Licence (British Columbia):** There

¹⁶English: <https://www.ontario.ca/page/ontarios-routine-immunization-schedule>, French: <https://www.ontario.ca/fr/page/calendrier-de-vaccination-systematique-de-lontario>

¹⁷French: <https://www.quebec.ca/transport/circulation-securite-routiere/periode-hivernale/pneus-hiver>, English: <https://www.quebec.ca/en/transport/traffic-road-safety/winter-road-safety/requirements-for-winter-tires>

¹⁸Both sources English-only: <https://www.icbc.com/driver-licensing/new-drivers/Graduated-licensing>, <https://www.icbc.com/driver-licensing/new-drivers/Get-your-L>

are 10 queries about the graduated licensing program for drivers in British Columbia.¹⁸ They cover topics such as the types of tests, age of eligibility, and rules for drivers with a learner's license. We note that recent changes to the Driver's Licence processes in British Columbia may result in some of these query-answer pairs becoming outdated.

- **Health Card (Alberta):** There are 10 queries related to the Alberta Health Care Insurance Plan (AHCIP) and eligibility.¹⁹ These are actu-

ally 5 distinct queries, where one uses the full name Alberta Health Care Insurance Plan and the other uses the acronym AHCIP, enabling us to examine how that difference in question formation may impact accuracy (future work).

- **Driving (New Brunswick):** There are 10 queries related to topics from the New Brunswick Driver Handbook.²⁰
- **Hunting (Nunavut):** There are 10 queries related to hunting and sport fishing in Nunavut.²¹ Several of these are yes/no questions.

¹⁹English: <https://www.alberta.ca/ahcip-eligibility>, <https://www.alberta.ca/ahcip-moving-to-alberta>, French: <https://www.alberta.ca/fr/ahcip-eligibility>, <https://www.alberta.ca/fr/ahcip-moving-to-alberta>

²⁰English: https://www2.gnb.ca/content/dam/gnb/Departments/ps-sp/pdf/drivers_vehicles/driverhandbook/part-4-traffic-on-the-highway-rules-of-the-road.pdf, https://www2.gnb.ca/content/dam/gnb/Departments/ps-sp/pdf/drivers_vehicles/driverhandbook/part-5-safe-driving.pdf, French: https://www2.gnb.ca/content/dam/gnb/Departments/ps-sp/pdf/drivers_vehicles/driverhandbook/partie-4-circulationregles-de-la-route.pdf, https://www2.gnb.ca/content/dam/gnb/Departments/ps-sp/pdf/drivers_vehicles/driverhandbook/partie-5-conduite-securitaire.pdf

²¹English: https://www.gov.nu.ca/sites/default/files/documents/2024-12/Nunavut_Hunting_Regulations_Guide_2024-25.pdf, <https://www.gov.nu.ca/sites/default/files/documents/2024-05/Nunavut%20Sport%20Fishing%20Guide%20Final%202024-2025.pdf>, French: (Not linked), <https://www.gov.nu.ca/sites/default/files/documents/2024-05/Nunavut%20Sport%20Fishing%20Guide%20Final%202024-2025%20-%20FRE.pdf>

Translators' Perceptions and Edit Traces: Quality of MT as a Tool in Canadian Parliamentary Translation

Jeniffer Leal-Wyss
Gabriel Bernier-Colborne
Delaney Lothian
Michel Simard
Rebecca Knowles

Jeniffer.Leal@nrc-cnrc.gc.ca
Gabriel.Bernier-Colborne@nrc-cnrc.gc.ca
Delaney.Lothian@nrc-cnrc.gc.ca
Michel.Simard@nrc-cnrc.gc.ca
Rebecca.Knowles@nrc-cnrc.gc.ca

Digital Technologies Research Centre, National Research Council of Canada, Ottawa, ON, Canada

Abstract

The Parliament of Canada's translation workflow includes access to a specialized neural machine translation (NMT) system. This study analyzes post-editing (PE) activity to identify the types of edits translators make when interacting with the NMT system, as well as the frequency, nature, and severity of errors encountered. We compare translations produced with and without the use of this NMT system to evaluate potential differences in edit patterns. To complement this analysis, we draw on insights from a user study. Our findings explore how translators' perceptions align with observed PE patterns and how their feedback can inform strategies to better understand, and possibly mitigate, some of the errors observed.

1 Introduction

At the Parliament of Canada, the transcripts of the debates in the House of Commons (and the Senate) are translated into both official languages, French and English. These translation services are provided by the Translation Bureau, the federal institution responsible for delivering translation services to the Government of Canada. As part of their workflow, translators for the House of Commons have the option to use a custom machine translation (MT) system called Hawkeye as a translation aid. Hawkeye is an NMT system integrated into their work environment based on the Sockeye toolkit (Hieber et al., 2022), trained on parliamentary text, and updated multiple times a year.

A previous study examining logs of user interactions with Hawkeye (Simard et al., 2026) in the House of Commons debates and committees texts showed an increase in MT usage, from 16% in May 2023 to close to 60% in December 2025. In that same work, a comparison of the number of changes

made by revisers (members of the parliamentary translation team who revise the translations before publication) to texts produced by translators with or without the help of Hawkeye showed that the texts produced with Hawkeye required a statistically significantly smaller number of reviser edits than texts produced without. While that work quantifies some differences in translations with and without this MT system, it doesn't provide us with information about the types of changes being made to the MT output or to the translators' reactions to MT errors. This motivates us to explore the impact of MT on translation and revision patterns in greater detail.

In this paper, we present a study that examines post-edited text samples, where we automatically detect spans that were modified during post-editing (PE) and revision, and then perform an analysis of error types, severity, and frequency using the MQM framework (Lommel et al., 2013). Additionally, we compare edits made during revision of texts translated with and without Hawkeye. Our analy-

sis shows that translations produced with the help of Hawkeye not only require fewer revisions, they also exhibit slightly lower error severity than those translated without Hawkeye. We complement this data with insights from a user study¹ involving parliamentary translators through questionnaires and interviews, highlighting their perceptions of MT quality. Through the analysis of text samples and users' perceptions, we gained valuable insights into the nature, severity and frequency of Hawkeye's errors, and its utility in this translation workflow. Our study supports that complementing data analysis with user feedback can contribute to a better understanding of translation quality.

2 Background and Related Work

In this study, we perform an assessment of MT output in a PE workflow. Thus, related research in translation studies and natural language processing provides valuable insights to inform this work.

The concept of quality in translation has often been debated, as it is inherently tied to expectations. However, research in translation quality evaluation (TQE) reminds us that quality should be defined in relation to the intended purpose of the translation or MT system being assessed (Bowker, 2020). In line with this perspective, we emphasize that the MT system under evaluation is designed to serve as an aid for translators, and will therefore be assessed according to that purpose.

Different methodologies and procedures have been used for TQE of MT, including human evaluation, which typically includes rating, ranking and annotation, as well as automatic methods, which are often based on lexical and/or semantic similarity to a reference translation. Human TQE is typically considered the gold standard, and it can be grouped into three main categories: manual scoring, semi-automatic, and task-based (Liu et al., 2024).

Semi-automatic evaluations are based on annotations from which a score is computed. For that purpose, a text is first annotated or post-edited, and then an edit distance algorithm can be used to identify edited spans in MT output (Snover et al., 2006). In cases where post-edited data are accessible, as in this study, this approach allows for analysis of corrections to MT output. This step can be complemented by various evaluation frameworks that

typically involve categorizing errors and assigning scores based on their number, type, and severity. One widely used framework is MQM (Lommel et al., 2013), which offers a fine-grained approach to TQE, although other valuable frameworks also exist (O'Brien, 2012; Graham et al., 2013; Kocmi et al., 2024; Magdy et al., 2026, i.a.). The use of a framework such as MQM can be valuable to understand error frequency, type, and severity, and to assess the general quality of a system, as shown by various studies (Ortiz-Boix and Matamala, 2017; Freitag et al., 2021; Park and Padó, 2024; Li et al., 2025). As MQM has a mostly standardized error taxonomy and is widely used for human evaluation, it can enable comparison of results across studies.

Other studies focusing on perceptions of MT in translation environments have employed more qualitative data collection methods. These perception-based methods are different from the ones described above in that they do not assess specific translation samples, but rather rely on subjective assessment, using instruments such as questionnaires, focus groups or interviews (Vieira and Alonso, 2020; Cadwell et al., 2018) or a combination of quantitative and qualitative data collection methods (Moorkens et al., 2018; Yang et al., 2023; Heinisch and Lušicky, 2019). Such research methods are helpful for understanding the usability of MT output for system users, capturing examples that might not be found in the samples selected for post-hoc evaluation by researchers, and assessing perceptions of MT quality as a computer-assisted translation (CAT) tool.

Closely related to the work described here are prior publications on MT and the Canadian Parliament, specifically the quantitative analysis of MT usage by translators over time in Simard et al. (2026) and the analysis of translators' reasons for adopting (or not adopting) the Hawkeye MT systems as part of their workflows described in Leal-Wyss et al. (2026). The latter draws on the the same pool of user study data that we analyze here, though it focuses on aspects of translator experience rather than questions of MT quality.

3 Methodology and Data

This paper uses two main approaches to assess quality and utility of MT: annotation and evaluation of two sets of translation samples (Sec. 3.1) and analy-

¹Study reviewed and approved by the National Research Council of Canada's Research Ethics Board (NRC-REB #2024-22).

sis of user perceptions of system quality (Sec. 3.2).

We were provided access to a large collection of Parliament translation log files containing the source text, the Hawkeye MT output, the initial translation (which may or may not have been based on Hawkeye output), the revised translation, a flag indicating whether the translator invoked Hawkeye to produce the initial draft, and other metadata (for more details, see Simard et al., 2026). We sampled data for annotation (Sec. 3.1.2) from this collection.

3.1 Annotation and Evaluation

We refer to the two subsets of our annotation and evaluation data as “revision data” (REV for short) and “post-editing and revision data” (PE+REV). For REV, we compare the translation drafts produced by translators to the final revised version; we split this into two groups: translation drafts produced with the use of Hawkeye or without the use of Hawkeye. For PE+REV, we compare the original Hawkeye MT output to the final revised version of the translation; this encompasses two steps: PE and revision together (hence the name).

3.1.1 Sample Annotation Schema

Two members of our team (fluent in English and French, with formal training in translation studies and experience in translation) performed blind MQM annotation on texts from both the REV and PE+REV datasets; these annotations were revised by the other annotator and disagreements were resolved through discussion. The PE+REV samples are a subset of the REV samples.

We automatically compared the first target text (Hawkeye MT in the case of PE+REV, translator’s draft in the case of REV) to the final revised version, highlighting and numbering the edited spans to make them clearly visible to the annotators. We then used the MQM framework (more specifically, MQM-Core²) for annotation, which organizes translation issues into top-level categories, such as *Accuracy* and *Terminology*, each of which includes subcategories. This is a subset of the full MQM schema comprising eight high-level error categories and the most common subtypes. It also includes a scoring model that rates errors by severity. In our sample analysis, we applied this framework to TQE of texts produced with and without the use of Hawkeye, then

calculated the frequency of error types and severity. Each highlighted error was assigned a top-level error type, subtype, and a severity level. In cases where multiple automatically-detected edits would be better interpreted as a single error, the annotators could cluster them into a single error. We do not compute global error scores weighted by severity, but rather use this analysis to understand edits made to MT output and to compare texts that used and didn’t use Hawkeye as a translation aid to assess which sample requires more edits and which sample shows more severe errors.

The annotation schema selected is not without challenges. One limitation is the absence of an explicit category for *preferential* changes. In such cases, we applied the severity level *Neutral* to indicate that these are considered stylistic preferences, and we classified them under *Style – Unidiomatic*. In many instances, this classification was chosen not because the initial draft was unidiomatic but because the revised version offered a more idiomatic rendering. We highlight that all errors classified as *Style – Unidiomatic* that have a *Neutral* severity are preferential – for those that were considered required changes, the *Minor* severity was used. We also found that certain errors could reasonably fit multiple categories, so the annotators discussed ambiguous errors and established their own guidelines to resolve these.

It is important to note that these are post-hoc analyses of edits; we do not have access to the actual tracked changes, which is why we use an edit distance algorithm to identify edited spans. It is possible that a translator has in fact deleted the MT output and then typed their own text from scratch (we do not have a way to automatically detect this), and we cannot assume that the non-Hawkeye subset of REV is MT-free, as translators may have post-edited the output of other (non-Hawkeye) MT systems.

3.1.2 Dataset for Annotation

From our collection of data, we sampled “chunks” (roughly paragraphs, which are further subdivided into “segments”, which are usually a single sentence, but some contain two or three) for which Hawkeye MT was used at translation time, uniformly at random; these are annotated in both the PE+REV and REV settings. We also sampled an

²As described in Lommel et al. (2024) and <https://themqm.org/the-mqm-typology/>

equal number of chunks for which Hawkeye MT was not used; these are annotated in only the REV setting. This produced a dataset containing 1393 segments. Statistics on this dataset are shown in Table 1. See Appendix A for details on data sampling; we focused only on House of Commons debates data for this annotation.

	PE+REV	REV
# chunks	114	230
# segments	784	1393
# words	14,737	26,197

Table 1: Numbers of chunks, segments, and words (i.e. space-separated tokens) in the sampled dataset.

3.2 User Study

The user study employed a mixed-methods explanatory sequential design (Creswell and Creswell, 2017). First, a questionnaire was used to collect data on users’ experience with Hawkeye.³ The findings from this stage informed the design of semi-structured interviews in the second stage, which were intended to expand upon the results of the questionnaire.

All participants for both stages were recruited from a pool of 66 parliamentary translators from the Translation Bureau. Both the questionnaire and interview protocol included consent forms, and participation was strictly voluntary. Participants were recruited via an email written by the research team and distributed by the Translation Bureau.

3.2.1 Questionnaire

The questionnaire was designed to gauge respondents’ experience with the Hawkeye MT tool. It was written in English and translated into French, and was deployed online using the Voxco survey software platform.

The questionnaire includes a matrix question regarding error types and frequencies, where participants were provided with MQM high-level categories along with a short explanation of each; a second matrix question regarding error severities and frequencies; a question regarding overall assessments of quality improvements over time; as well as a question about the error types that users would most like to have eliminated. For the matrix ques-

tion, participants were presented with the eight top-level MQM-Core categories and asked to indicate the frequency of each type of error. We chose not to include the subcategories, as these would have been difficult to assess based solely on impressions and memory without reference to specific translation samples. This design aimed to capture general perceptions of the frequency of error types in the MT output. We intentionally kept error types and frequency distinct from severity, using two separate matrix questions, since integrated designs are more appropriate when evaluating specific translation samples. Treating them separately allowed us to assess these dimensions as two complementary but independent aspects of MT performance.

We sent 66 questionnaire invitations and received 39 responses (total response rate: 59%). Of these, 4 responses were excluded after determining that participants were confused about the tool under evaluation. In this paper, we present only the section of the questionnaire intended for participants who had used Hawkeye (a subset of the original 35 valid responses, which is $n=26$), since they are the only participants who can assess the system’s quality. We analyzed the quantitative data using descriptive statistics and used thematic analysis (Naeem et al., 2023) for the open-ended questions.

3.2.2 Interviews

The second stage of this user study used semi-structured interviews, which asked participants about their perception of the Hawkeye MT system and other CAT tools. Invitation e-mails were sent to participants who had volunteered through an optional form at the end of the questionnaire. Interviews were held remotely in French or English, led by a primary interviewer with support from a notetaker. They lasted up to 45 minutes and were all recorded. In total, 10 participants were interviewed. Thematic analysis used an inductive coding process, in which themes were identified and collected from participant open answer data.

4 Analysis of Edits

Overall error counts and frequencies in both the REV and PE+REV data are shown in Table 2. An error in this context is a manually annotated error (with MQM category and severity label) which

³For the full text of the questionnaire, see the appendices of Leal-Wyss et al. (2026).

	REV-ALL		REV-NO-HMT		REV-WITH-HMT		PE+REV	
	EN-FR	FR-EN	EN-FR	FR-EN	EN-FR	FR-EN	EN-FR	FR-EN
# errors	379	381	206	194	173	187	631	641
# segments	668	725	282	336	386	389	398	386
Errors/seg.	0.567	0.526	0.730	0.577	0.448	0.481	1.585	1.661

Table 2: Error frequencies. “HMT” means Hawkeye Machine Translation. An error (manually annotated) can correspond to one or more of the automatically-detected edit spans (merged by annotators when multiple automatically-detected errors spans could best be understood as a single error).

can correspond to one or more of the automatically-detected edit spans. The frequency distribution of error counts per segment is shown in Appendix C. While PE+REV counts are shown in the same table, they should not be directly compared to the REV counts, as the former covers the full set of changes from raw MT output to revision, while the latter covers only changes from the first translator’s draft to revision.

4.1 Edits when Hawkeye is Used

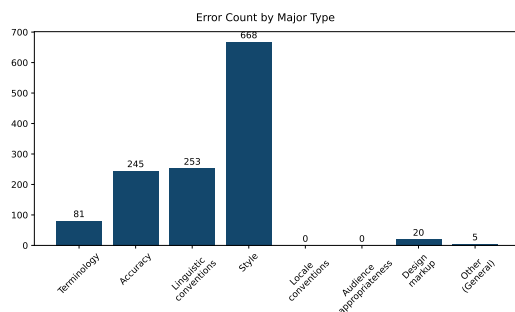


Figure 1: Frequency distribution of error categories in the PE+REV data

In this section, we present our results related to edits made to the Hawkeye MT output in PE+REV with a focus on two aspects: error type and severity. As shown in Fig. 1 and Fig. 2, the four most frequent error categories identified in PE+REV are *style*, *linguistic conventions*, *accuracy* and *terminology*, while the two most frequent error severity categories identified are *Minor* and *Neutral* errors. Note that we obtained similar distributions for the REV data (see Fig. 8 in Appendix D and Fig. 9 in Appendix E).

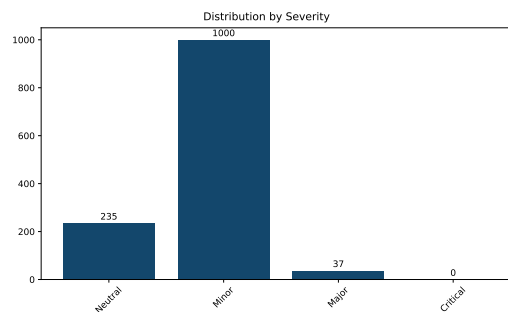


Figure 2: Frequency distribution of error severities in the PE+REV data

In Figure 3, we present a Sankey diagram showing the main categories and subcategories of errors in the PE+REV data along with their corresponding severities (the analogous diagram for the REV data is shown in Fig. 12, Appendix G).

A closer examination of style errors reveals that many of these corrections relate to unidiomatic or awkward phrasing, as well as non-compliance with organizational style. The latter occur when the MT output does not adhere to institutional guidelines, for example, using regular spaces instead of non-breaking spaces, regular hyphens instead of non-breaking hyphens, or quotation marks that differ from those recommended in internal style guidelines, among others.

Accuracy errors most often involve mistranslations or omissions. In the case of mistranslations, we observed recurring issues with false friends and overly literal renderings. Although these issues may appear quite different from a style error that results in unidiomatic phrasing, these behaviours are not entirely unrelated in terms of their explainability, as they are both caused by a translation that is too close to the source text. Moreover, in the context of MT, it may be preferable for a system to produce a trans-

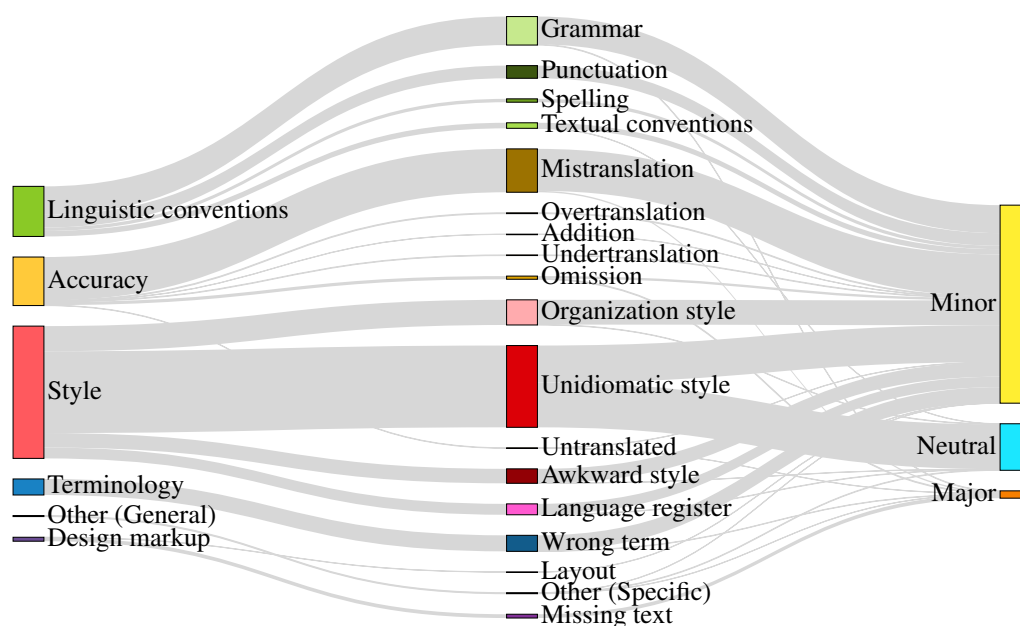


Figure 3: Sankey plot of Error Categories, Subcategories, and Severities in the PE+REV data.

lation that remains closer to the source text than one that introduces hallucinations.

Errors related to terminology corresponded in all instances to the use of an incorrect term (e.g. “bills” being translated as “*projets de loi*” instead of “*factures*”). Terminology issues were not necessarily tied to parliamentary language, and involved a variety of possible explanations, including the training data and literal translations (false friends).

For linguistic convention issues, we found that these most often involved grammar, textual conventions (e.g. cohesion, coherence, repetition), spelling, and punctuation issues. In many instances, errors occurred in the use of a tense, gender, or number within a given verb form, which could be attributed to the MT system’s limited context scope (individual segments). We often noticed cases of wrong gender, especially when translating into French, (for example, “*le député*” instead of “*la députée*”). These patterns suggest that these errors are in many cases related to the fact that the system operates at the sentence level and therefore lacks access to contextual information beyond individual sentences, or world knowledge.

Our analysis of PE+REV samples did not find any issues related to locale conventions (such as

number, currency, measurement, date or time format), or issues related to audience appropriateness (which can include culture-specific references or offensive language). We suspect this to be related to the parliamentary data used for training.

As noted above, most errors were found to be *minor* or *neutral* in severity, with only a few identified as *major* and none as *critical*. This distribution reflects the desired behaviour for an MT system intended for PE, where corrections involve non-critical issues that, if left unedited, would have minimal impact on the target audience. Thus, the MT output quality is deemed as adequate for PE in a human-centred process where the MT tool serves as an aid for translators.

Regarding the mapping between edited spans and discrete translation errors, the distribution of the number of spans per error (shown in Appendix F) indicates that the vast majority of errors map to a single edited span, but some errors map to 2 or 3 spans. In rare cases (likely, preferential stylistic edits that we merged together), the number of edits is even higher.

4.2 Impact of Hawkeye MT on Revision

We observe that REV data (revisions of translator drafts) for which Hawkeye was used in the initial draft contained fewer errors than segments for which it was not, as shown in Table 2.

REV samples show 0.567 errors per segment EN-FR and 0.526 FR-EN, on average. Of these, texts translated using Hawkeye exhibited 0.448 errors per segment in EN-FR and 0.481 in FR-EN, while texts translated without Hawkeye (REV-NO-HMT) showed a higher number of errors per segment in EN-FR (0.730) than in FR-EN (0.577). This aligns with the results of the initial log file analysis that text translated with the help of Hawkeye required less revision than text translated without it. The difference in errors/segment between REV-NO-HMT and REV-WITH-HMT is statistically significant only in the EN-FR direction.⁴

The distribution of error severity did not show notable differences between texts translated with Hawkeye and texts translated without Hawkeye. In both cases, a majority of errors were found to be either *minor* or *neutral*, with very few *major* errors (see Fig. 9 in Appendix E). We observe, however, that the only critical error found in our sample comes from a text that did not use Hawkeye, corresponding to a sentence that was entirely omitted, and that the number of major errors is higher in texts that didn't use Hawkeye, suggesting that its use can contribute to quality improvements.

In terms of error types, the texts translated with and without Hawkeye followed similar trends, though with some variation in distribution. The most common error category for translations produced with Hawkeye was style errors, followed by issues related to linguistic conventions, accuracy, terminology, and design/markup. Similarly, translations produced without Hawkeye also showed a predominance of style errors, followed by linguistic conventions, accuracy, terminology, and design/markup.

5 User Perceptions of MT Quality

Our questionnaire aimed to identify the types of errors translators perceived to be present in the MT output. To do so, the 26 participants who had used Hawkeye were presented with a matrix question based on MQM error categories, with frequency op-

tions. Based on the responses, summarized in Fig. 4, the most frequent MT errors are perceived to be related to *Accuracy* with 23 and *Terminology* with 21 responses indicating that these are perceived to occur sometimes or often, followed by *Style* errors, with 14 respondents reporting them sometimes or often. *Linguistic conventions* errors are perceived to occur less frequently, with 9 responses noting that these appear sometimes or often, followed by *audience appropriateness* with 4 participants reporting them as sometimes or often, while *design and markup* errors are reported to appear sometimes or often by 5 participants, and *locale conventions* errors are reported to be less frequent with 3 participants noting these appear sometimes or often.

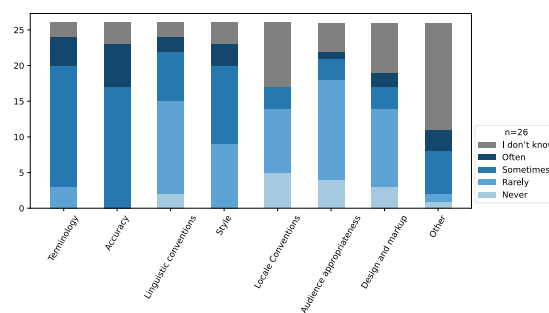


Figure 4: Question: *When using the Hawkeye translation tool, what types of errors do you usually observe, and how frequently do you observe them?*

We also asked these 26 participants a matrix question about error severity to determine how often *critical*, *major*, *minor*, and *neutral* errors were perceived to be found in the NMT output. Responses are summarized in Fig. 5. The most frequently reported severity was *minor errors* with 22 participants reporting these appeared sometimes or often, followed by *neutral errors* with 21 reporting them sometimes or often. Conversely, *major errors* were reported to be less frequent, with 13 participants reporting they appear sometimes and no participants reporting they appear often, and *critical errors* were the least frequent with 2 participants reporting them sometimes and no participants reporting them often. This suggests that less severe errors are perceived to occur more frequently, whereas more severe errors are relatively rare.

⁴Mann-Whitney U test (since the distribution is not normal) with $p < 0.01$.

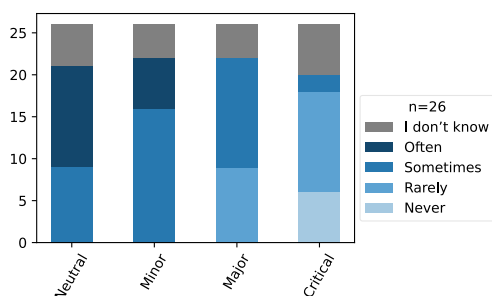


Figure 5: Question: *Based on the above defined error severities, please indicate the frequency of error severity occurrence in the Hawkeye translation output.*

We also asked these 26 participants about the types of errors they would most like to see eliminated from the MT output (multiple responses allowed). The three most frequent answers were accuracy (23 responses), terminology (17), and linguistic conventions (7). They also had the option to select “Other” and provide comments in an open text box. Here, translators noted that they often modify the MT output because it is too literal or too close to the source text. Others highlighted issues such as inconsistent terminology, some hallucinations involving proper names, reduced lexical variation, and stylistic problems specific to the House of Commons.

During the interviews, we explored quality in greater depth. We asked participants (n=10) about the types of errors they observed in the MT system, and how these compared to human errors (*Consider the types of errors that you observe in the Hawkeye machine translation output. How do they compare against errors you might have observed in human translation?*). Participants mentioned issues commonly associated with MT systems, including challenges with informal or everyday language, idiomatic expressions, figurative language, literal formulations, long and complex paragraphs, co-reference (difficulties identifying correct pronouns), proper names, omissions, hallucinations, additions, as well as misguided attempts to disambiguate vague discourse. Many interviewees also noted that the system’s output was generally accurate when handling specialized parliamentary language (6/10 participants) but tended to perform less effectively with more general language—a differ-

ence likely related to the data used to train the system. We illustrate this issue with an example taken from the interviews:

“I remember a specific example: someone was thanking his wife for taking care of the cat litter on her own, and Hawkeye rendered it as something like a ‘trash can’ for the cats or something along those lines, even though it’s really not a specialized or difficult term to translate.” (English translation, Interview Participant 4)⁵

This set of challenges for MT have been previously discussed in the literature (Naveen and Trojovský, 2024) and some are known limitations of NMT systems. While some of these issues may be difficult to fully resolve, others may be easier to address. For instance, co-reference and consistency issues might be possible to improve on through the use of larger context windows or with external information sources. Overall, even considering the observations about MT output quality, it appears that translators perceived the quality to be acceptable in a PE setting, as most errors are reported to be either minor or neutral, and only a minority are perceived to be critical or major.

6 Comparison Between Perceptions and What Translators Change

While we have found differences between actual post-edits observed in our sample and perceived MT errors reported by translators, we have also found convergence in other areas. For example, in terms of error severity, most errors in our sample analysis were found to be *minor* or *neutral*, with few *major* errors, which matches participants’ perceptions as reported in our user study.

Regarding perception of error types reported in our user study, we observed slight differences compared to the annotated samples, but overall trends remained similar. In our sample, the most common post-edits fell within the categories of *style*, *linguistic conventions*, *accuracy*, and *terminology*. This finding is noteworthy, as these categories were also perceived as frequently found in the MT system and as those which translators would most like to see eliminated. However, the distribution shows slight variations. One source of these differences could be that user study participants are describing all of their experiences with Hawkeye (both debates and com-

⁵Original French versions of quotations are provided in Appendix B.

mittees), while our annotations only focus on debates data; in particular, it may be the case that some committees use more challenging technical terminology. These differences could also be explained by participants' varying levels of familiarity with the MQM framework, despite the questionnaire including a brief explanation of its categories, or because some MQM categories may be particularly open to variable interpretations. Other aspects, such as how memorable errors are, or how challenging they are to revise, may also play a role in this discrepancy and in the similarity between the distribution of the errors perceived as frequent and those that translators would most like to see improved.

Errors related to *linguistic conventions* were the second most common errors in our annotations, while participants perceived them as less common than terminology, accuracy, or style errors. The most common of these errors observed were related mainly to grammar, including changes in tense, gender or number agreement, pronouns, and prepositions. Although these errors were not perceived as particularly frequent in the questionnaire responses, feedback collected during the interviews did identify them as common issues in Hawkeye's output. This finding aligns with our sample analysis, as illustrated by one participant: "*Certainly, one of the most common errors that I pick up in Hawkeye is... and you know, I translate into English, I don't think this is as big a problem going into French, they have their own set of problems, but... Using 'he' or 'she' instead of 'it'. That happens a lot in Hawkeye.*" (English original quotation, Interview participant 2).

Additionally, *terminology* errors were perceived to be among the most common error types, while our annotation sample found them to be less common. This may reflect the fact that certain issues, such as the incorrect translation of a term, could have been interpreted as mistranslations, which fall under the category of *accuracy*. For example, in our sample, the translation of "food" as "nourriture" was corrected to "denrées alimentaires." This example shows a case that could cause confusion because "food" may not necessarily be considered a term in most contexts, although it was in this specific context. This example highlights a limitation of this framework, as well as of others: certain nuances may rely on personal interpretation.

The most common errors found in our samples

were *style*-related. Although translators also perceived *style* issues as frequent, they did not consider them the most common error type in Hawkeye's output. Certain client-required revisions, categorized as *organizational style* in our analysis, may have been classified under different categories by translators. For example, this could occur when corrections are made to comply with capitalization rules that are specific to the House of Commons. While capitalization issues would normally be considered *linguistic convention* errors in many cases, certain capitalization corrections would be acceptable in other contexts. Because the correction is made to ensure compliance with the client's style guide, the error is more appropriately categorized as a *Style – Organizational Style* error.

Regarding *accuracy* errors, translators identified these as one of the most frequent categories in Hawkeye's output. They noted frequent mistranslations of idiomatic expressions and overly literal translations, as well as occasional omissions and hallucinations, particularly when the MT attempted to disambiguate vague discourse. This is consistent with the findings from our sample analysis, in which *mistranslation* was the most common subcategory of *accuracy* issues.

Additionally, the types of errors that translators reported as infrequent, such as *audience appropriateness*, which can include cultural-specific references or offensive language, and *locale conventions*, which can include number, currency, and measurement formats, among others, were not identified in our sample analysis. This is likely because the translators worked with greater volumes of texts, where such issues may arise occasionally, but they did not appear in our relatively small sample.

An interesting finding from our analysis is that many of the edits made by translators are closely tied to the text type. Parliamentary translators often work with transcripts of speeches, which are naturally vague, ambiguous, or loosely structured due to the characteristics of spoken language. As a result, translators frequently have to restructure sentences or adapt them for the target audience, and add information or cues that are implicit in the context of the speech but not explicitly stated by the speaker, helping to clarify meaning for the reader, or otherwise adapt the text to the target audience. Currently, most MT systems are not capable of performing this

kind of contextual enrichment, making human intervention necessary. A comment from one of the interviewees supports this observation: *“To be honest, sometimes there are things that just can’t be translated—or that translate extremely poorly—and at that point, it’s no longer a matter of translation; it’s outright adaptation, so... and I don’t think that’s what we’re asking from the tool for now.”* (English translation, Interview Participant 3)

We would like to highlight that the feedback provided by translators during our user study complements the data collected through our sample analysis. The insights gathered offer valuable examples and explanations about parliamentary translation that are essential for understanding Hawkeye’s quality and could help guide future work with the system. Moreover, the feedback sheds light on recurring issues and potential areas for improvement.

7 Limitations and Future Work

A key limitation of this study concerns the sample evaluation. Although our research team made use of all available user reference materials and engaged annotators with expertise in translation, the fact that the annotators were not the actual users introduces the possibility of judgment discrepancies. Users may categorize certain errors differently, drawing on their own knowledge and reference materials that were not accessible to our team. Despite this limitation, we consider the evaluation framework to be valuable, as it provides a meaningful overview of the types of errors identified in the MT system under study. Further, collecting multiple, independent annotations would allow us to measure agreement and identify the most confusable pairs of error categories, and possibly revise our annotation schema.

While MQM analysis proved valuable for comparing perceived errors to actual PE, we acknowledge that this framework has limitations in this context. Identifying error types and assessing their severity provide meaningful insights into a system’s overall performance and behaviour; however, to achieve a deeper understanding of the factors driving specific system behaviours, an additional dimension—error explainability—is needed. Going forward, we could perform a targeted analysis of error samples that considers error type, explanation of causes, and potential resolution. For example, if a terminology error is traced back to training data, an

appropriate remedy could be to augment the training corpus with additional relevant data.

For convenience, we have treated a sequence of MT systems deployed overtime as a single system. Since the user study was conducted only once, it does not allow for a fine-grained analysis of changes in MT output over time. However, responses to a question on the evolution of system performance showed that a plurality of Hawkeye users perceived an improvement in MT output quality. Further information on the system’s evolution can be found in [Simard et al. \(2026\)](#), which examines the evolution of system adoption and compares the number of revisions made to Hawkeye and non-Hawkeye translations over time. We also note that the main change in the updated systems during the period covered by this study was the use of increasing amounts of training data.

Finally, we note that in previous work ([Simard et al., 2026](#)), we found that most translators used Hawkeye either rarely or almost all the time. As a result, when comparing translations done with and without the help of Hawkeye, we are in fact comparing translations done by two mostly distinct groups of translators. This means that within our current study, we cannot fully disentangle differences based on the use of Hawkeye from individual translator differences.

8 Conclusions

Our study provides evidence that Hawkeye is a useful tool for translators working for the Canadian Parliament, as its use leads to a reduced number of edits by revisers compared to translations produced without the Hawkeye system, and to less severe errors. We note that our comparisons are observational; however, the insights gained from this study could help improve the system and, in turn, have a direct impact on translators’ working conditions.

This study also provided valuable insight into how frequently certain types of errors occur and which of them may be more easily addressed. Such understanding would not have been possible through sample analysis alone; it was enriched by the feedback and perspectives of translators who regularly interact with the system. For this reason, we believe that close collaboration between developers and users is essential to ensure a process of continuous improvement of NMT systems.

Acknowledgements

We would like to thank our study participants, the translators working for the parliamentary translation team, as well as the Canadian government's Translation Bureau, without whom this work would not have been possible. We also thank Kerry Butt for his assistance with the deployment of our questionnaire. Lastly, we would like to thank our colleagues Marc Tessier and Jackie Lo for feedback on this work.

References

- Bowker, L. (2020). Fit-for-purpose translation. In O'Hagan, M., editor, *The Routledge Handbook of Translation and Technology*, volume 1, pages 453–468. Routledge, 1st edition.
- Cadwell, P., O'Brien, S., and Teixeira, C. (2018). Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators. *Perspectives*, pages 1–21.
- Creswell, J. W. and Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.
- Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., and Macherey, W. (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.
- Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2013). Continuous measurement scales in human evaluation of machine translation. In Pareja-Lora, A., Liakata, M., and Dipper, S., editors, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.
- Heinisch, B. and Lušický, V. (2019). User expectations towards machine translation: A case study. In Forcada, M., Way, A., Tinsley, J., Shterionov, D., Rico, C., and Gaspari, F., editors, *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 42–48, Dublin, Ireland. European Association for Machine Translation.
- Hieber, F., Denzumi, C. D., Niu, X., Hoang, C., Tran, K., Hsu, B., Nadejde, M., Lakew, S., Mathur, P., Currey, A., and Federico, M. (2022). Sockeye 3: Fast neural machine translation with pytorch.
- Kocmi, T., Zouhar, V., Avramidis, E., Grundkiewicz, R., Karpinska, M., Popović, M., Sachan, M., and Shmatova, M. (2024). Error span annotation: A balanced approach for human evaluation of machine translation. In Haddow, B., Kocmi, T., Koehn, P., and Monz, C., editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.
- Leal-Wyss, J., Lothian, D., Bernier-Colborne, G., Knowles, R., and Simard, M. (2026). “all those in favour will please say yea”: Understanding the factors behind machine translation adoption at the Canadian Parliament. In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026)*, pages 593–618, Tilburg, The Netherlands. European Association for Machine Translation.
- Li, Y., Suzuki, J., Morishita, M., Abe, K., and Inui, K. (2025). MQM-chat: Multidimensional quality metrics for chat translation. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3283–3299, Abu Dhabi, UAE. Association for Computational Linguistics.
- Liu, T., Lo, C.-k., Marshman, E., and Knowles, R. (2024). Evaluation briefs: Drawing on translation studies for human evaluation of MT. In Knowles, R., Eriguchi, A., and Goel, S., editors, *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 190–208, Chicago, USA. Association for Machine Translation in the Americas.
- Lommel, A., Gladkoff, S., Melby, A., Wright, S. E., Strandvik, I., Gasova, K., Vaasa, A., Benzo, A., Marazato Sparano, R., Foresi, M., Innis, J., Han, L., and Nenadic, G. (2024). The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. In Martindale, M., Campbell, J., Savenkov, K., and Goel, S., editors, *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas*, pages 75–94, Chicago, USA. Association for Machine Translation in the Americas.

- Lommel, A. R., Burchardt, A., and Uszkoreit, H. (2013). Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.
- Magdy, S. M., Alwajih, F., Mekki, A. E., El-Sayed, W., and Abdul-Mageed, M. (2026). Lqm: Linguistically motivated multidimensional quality metrics for machine translation.
- Moorkens, J., Toral, A., Castilho, S., and Way, A. (2018). Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7:240–262.
- Naeem, M., Ozuem, W., Howell, K., and Ranfagni, S. (2023). A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. *International Journal of Qualitative Methods*, 22:1–18.
- Naveen, P. and Trojovský, P. (2024). Overview and challenges of machine translation for contextually appropriate translations. *iScience*, 27(10):110878.
- O'Brien, S. (2012). Towards a dynamic quality evaluation model for translation. *The Journal of Specialised Translation*, pages 55–77.
- Ortiz-Boix, C. and Matamala, A. (2017). Assessing the quality of post-edited wildlife documentaries. *Perspectives*, 25:1–23.
- Park, D. and Padó, S. (2024). Multi-dimensional machine translation evaluation: Model evaluation and resource for Korean. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11723–11744, Torino, Italia. ELRA and ICCL.
- Simard, M., Leal-Wyss, J., Bernier-Colborne, G., and Knowles, R. a. (2026). A longitudinal study of the adoption of specialized mt systems in Canadian Parliamentary translation. In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation (EAMT 2026)*, pages 133–141, Tilburg, The Netherlands. European Association for Machine Translation.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (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.
- Vieira, L. N. and Alonso, E. (2020). Translating perceptions and managing expectations: an analysis of management and production perspectives on machine translation. *Perspectives: Studies in Translatology*, 28(2):163–184.
- Yang, Y., Liu, R., Qian, X., and Ni, J. (2023). Performance and perception: machine translation post-editing in chinese-english news translation by novice translators. *Humanities & Social Sciences Communications*, 10(1):Article 798.

A Details on Data Sampling

We sampled data for annotation separately from two pools: data from 2023-2024, which ranges from 2023-05-15 to 2024-12-17 and data from 2025, which ranges from 2025-05-27 to 2025-06-20. While the full 2025 data available to us covered a wider span of time (from 2025-04-30 to 2025-12-12), there was very little data in April (explaining why our random sample does not cover that month) and we intentionally set a cutoff date of 2025-07-01 to make sure that we were only looking at data that translators had worked with prior to the start of our questionnaire. Gaps in the data can be accounted for by the fact that we focused only on translation for the debates of the House of Commons, which are translated almost exclusively while the House is sitting.

We sampled “chunks” for which Hawkeye MT was used at translation time from each of the datasets, uniformly at random. The size of the sample was constrained by the time it takes to complete the manual annotations; our annotated REV data comprises approximately 0.10% of the 77091 chunks in the 2025 data and 0.07% of the 209767 chunks in the 2023-2024 data, biasing our annotations slightly toward the data closest temporally to the questionnaire. Statistics for the two date ranges are shown in Table 3.

	PE+REV		REV	
	Chunks	Segs.	Chunks	Segs.
2023-2024	74	583	150	1024
2025	40	201	80	369
Total	114	784	230	1393

Table 3: Numbers of chunks and segments in the sampled dataset.

We intended to have equal numbers of REV-WITH-HMT and PE+REV segments (Table 2), but they are slightly different. These differences are due to the fact that (1) we accidentally missed one of the 115 chunks when producing the PE+REV dataset, as shown in Table 1; and (2) there were slight differences in the segmentation of chunks into segments, between the REV and PE+REV data.

B Original Interview Quotations and Notes on User Study

Here we provide the original French transcriptions of the two translated quotations. Translation was performed by a member of the research team with a background in French–English translation.

“Je me souviens un exemple précis, il y a quelqu’un qui remercie son épouse de s’occuper de la litière des chats toute seule, puis Hawkeye m’avait sorti comme une ‘boîte à déchets’ pour les chats ou quelque chose comme ça, alors que ce n’est vraiment pas un terme spécialisé, compliqué à traduire”. (Original French Quotation, Participant 4)

“En même temps, si on est franc, parfois, il y a des trucs qui ne se traduisent pas ou qui se traduisent extrêmement mal, puis là on n’est plus dans la traduction, on est carrément dans l’adaptation, donc... et je ne pense pas que c’est ce qu’on demande à l’outil pour l’instant”. (Original French Quotation, Participant 3)

For the full text of the questionnaire and the initial interview questions, see Appendix B of [Leal-Wyss et al. \(2026\)](#).

C Distribution of Errors/Segment

The frequency distribution of error counts per segment in the PE+REV data is shown in Fig. 6.

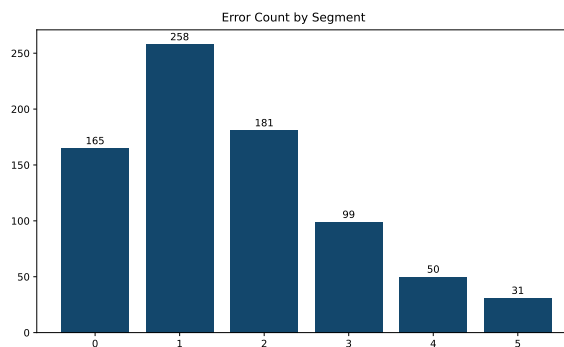


Figure 6: Frequency distribution of error counts per segment in the PE+REV data.

The frequency distribution of error counts per segment in the REV data is shown in Fig. 7.

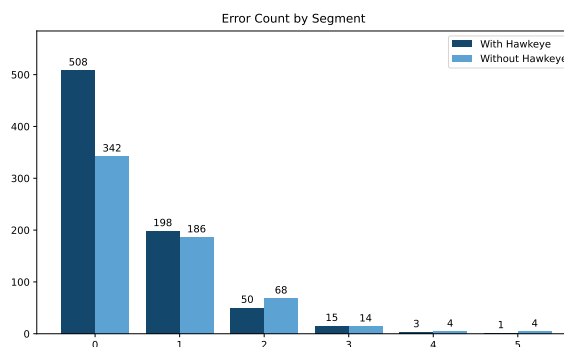


Figure 7: Frequency distribution of error counts per segment in the REV data.

D Distribution of Error Categories

The frequency distribution of error categories in the REV data is shown in Fig. 8.

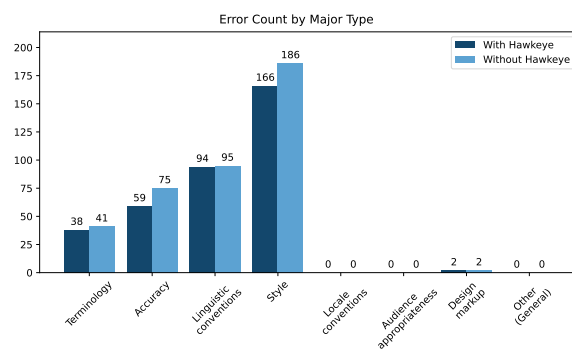


Figure 8: Frequency distribution of error categories in the REV data.

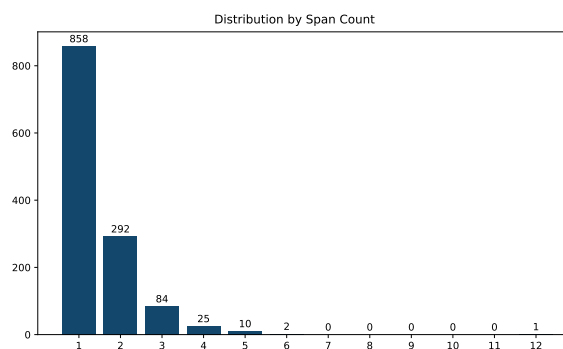


Figure 10: Frequency distribution of span count per error in the PE+REV data.

E Distribution of Error Severities

The frequency distribution of error severities in the REV data is shown in Fig. 9.

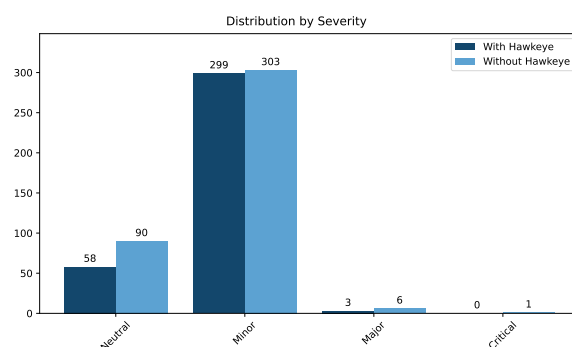


Figure 9: Frequency distribution of error severities in the REV data.

The frequency distribution of span count per error in the REV data is shown in Fig. 11.

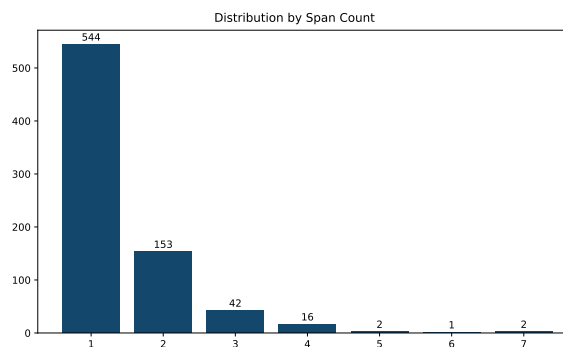


Figure 11: Frequency distribution of span count per error in the REV data.

F Distribution of Spans/Error

The frequency distribution of span count per error in the PE+REV data is shown in Fig. 10.

G Relationship Between Error Categories, Subcategories, and Severities

The relationship between error categories, subcategories, and severities is shown in Fig. 12 for the REV data (the analogous figure for PE+REV was shown in Fig. 3).

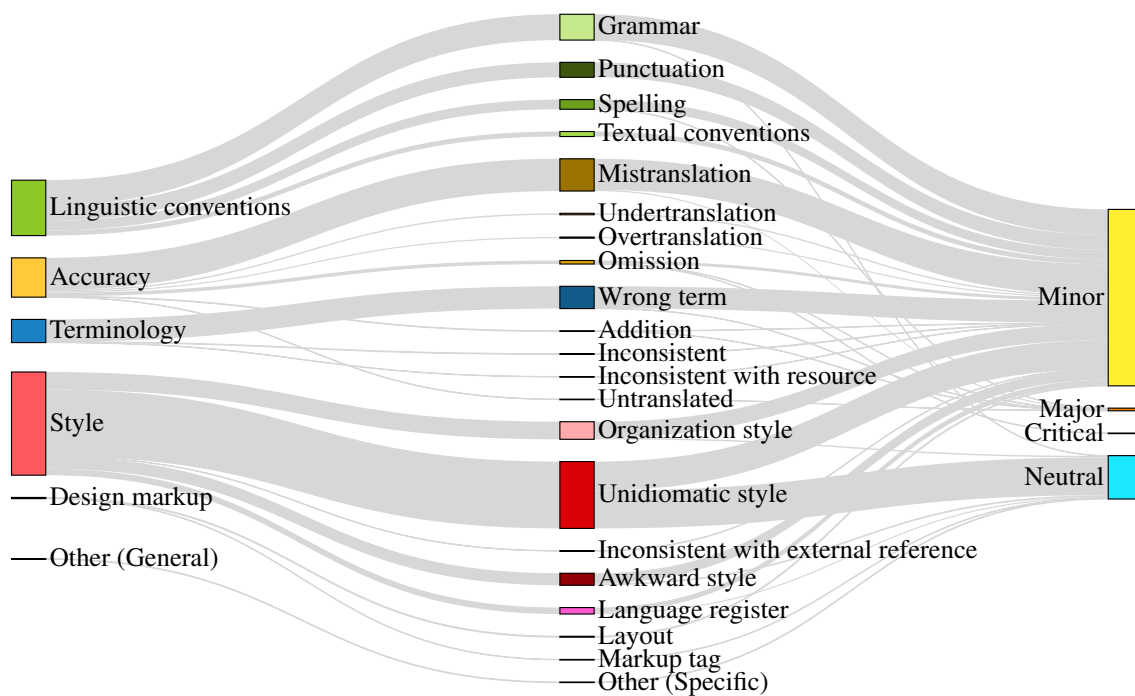


Figure 12: Error Categories, Subcategories, and Severities in the REV data.

Document Summarization for AI-based Post-Editing

Vera Senderowicz Guerra
Dimitrios Pavlou
Peter Bourgonje
Olesia Khrapunova
Konstantinos Karageorgos
Aaron Schliem
Welocalize

vera.senderowicz@welocalize.com
dimitrios.pavlou@welocalize.com
peter.bourgonje@welocalize.com
olesia.khrapunova@welocalize.com
konstantinos.karageo@welocalize.com
aaron.schliem@welocalize.com

Abstract

Post-Editing (PE) is typically performed on isolated segments or small batches, without access to broader document context. In this paper, we investigate whether pre-generated, document-level summaries can improve PE quality. Using a purpose-built summarization prompt evaluated across nine LLMs from OpenAI and Google, we select two models with contrasting summary styles for downstream experiments on 448 documents covering 37 target locales and 13 content domains. Summaries generated by `gemin-2.5-flash-lite`, which are directive and domain-specific, bring output closer to the human reference and yields modest gains in COMET, whereas those generated by `GPT-4o`, which tend to be more generic and descriptive, degrade performance across most metrics. The positive effect appears most pronounced in terminologically dense domains and lower-resource locales. A qualitative analysis shows that improvements arise when summaries provide specific, actionable guidance on terminology, domain conventions, and style, and that performance decreases when summaries are underspecified or conflicting. These findings suggest that summary *specificity and actionability*, rather than the mere addition of context, determine whether document-level information benefits post-editing.

1 Introduction

A growing body of work demonstrates that AI Post-Editing (AIPE)¹ enhances Machine Translation (MT) output and Translation Memory (TM) fuzzy matches (Koehn and Senellart, 2010; Raunak et al., 2023; Uguet et al., 2024). This effectiveness is further amplified when prompted with language-specific conventions and critically selected historical human-approved translations (Moslem et al., 2023; Mu et al., 2023; Chitale et al., 2024). Building on this evidence, we hypothesize that AIPE further benefits from document-relevant context, and we experiment with providing the PE model with a pre-generated, high-level summary of the entire source document, including its content type and domain.

Our contributions are as follows:

- We propose augmenting AIPE prompts with automatically generated document-level summaries, giving the model access to global context when making segment-level post-editing decisions.
- We describe a summary generation pipeline in which we first generate summaries across multiple LLMs, use human evaluation to identify the target quality and tone, recover a reusable prompt via cross-model reverse prompt engineering, and then shortlist two models with contrasting summary styles for downstream evaluation.

¹Throughout this paper, AIPE refers to the automatic editing of MT output or TM matches by an LLM, as distinct from MT post-editing (MTPE), in which a human editor revises MT output.

- We evaluate our approach on 448 documents covering 37 target locales and 13 domains. We find that summaries improve PE when they provide specific, actionable guidance on terminology, domain conventions, and style, but degrade it when underspecified or conflicting. The effect is further modulated by content domain and target language.

2 Related Work

Translating segments in isolation leads to discourse-level errors such as inconsistent terminology and pronoun misalignment (Voita et al., 2019). While document-level MT approaches address this by providing broader context, Herold and Ney (2023) show that volume alone does not help and can introduce noise, motivating *distilled* over raw context. Mohammed and Niculae (2025) further demonstrate that Large Language Models (LLMs) are sensitive to whether context is genuinely informative, and uninformative context can harm performance. Together, these findings argue in favor of providing a structured summary rather than adjacent segments.

With respect to PE, Moslem et al. (2023) show that in-context signals such as style guides and fuzzy matches enable LLMs to outperform dedicated MT engines on domain-specific content without fine-tuning. Our AIPE baseline is based on this LLM-based paradigm. Ki and Carpuat (2024) establish that signal *specificity* is decisive: Fine-grained feedback improves PE quality, while coarse signals do not. This aligns with our findings that more directive summaries improve but more generic summaries decrease performance.

3 Method

Our set-up consists of the two main components described below: the generation of document-specific summaries (Section 3.1), and subsequently using these tailored summaries during PE (Section 3.2).

3.1 Summary Generation

We used two non-reasoning (GPT-4o, GPT-4.1) and two reasoning (o1, o3) models through the OpenAI GUI² to iteratively generate summaries for two English source documents. From our pool of documents (see Section 4), we randomly selected one

from the Marketing domain and another consisting of Technical Documentation, to cover diverse document genres. For each document, all source segments were concatenated into a single input text and passed to the summary generation model. We shared the output summaries, along with the original source documents, with a team of professional linguists, and asked them to rank the summaries by their perceived usefulness in the context of a translation/post-editing task. Overall agreement favoured the summaries generated by the o1 model. We then reverse-prompted GPT-4o with the preferred summary, instead of o1, to avoid model-specific bias in the recovered prompt, and obtain a generation instruction that is not overfit to a single model. Li and Klabjan (2025) demonstrate that reverse prompt engineering yields semantically faithful results even across different generator-inverter model pairs, supporting the feasibility of this cross-model approach. The final prompt to create the summaries is included in Appendix B.

This prompt was then used to automatically produce summaries for a wider set of documents representing all content types present in our evaluation data (see Section 4). At this stage, we experimented with several OpenAI and Google models³, spanning a range of cost, latency, and reasoning profiles. To determine their suitability for use in AIPE prompting, we had the resulting summaries qualitatively evaluated by a panel of linguists and engineers, who ranked them according to post-editing utility and usability in prompting contexts, respectively. Appendix A describes the panel’s assessment and verdict reasons in more detail, but the final choice was to keep one model per provider: GPT-4o and gemini-2.5-flash-lite.

3.2 Post-Editing

Our AIPE system operates on both MT output and TM fuzzy matches, leveraging publicly available LLMs prompted with the target-locale-specific style guide and in-context source–target pairs retrieved from previously approved projects or translation memories. In our experiments, we augment this baseline by prepending to the prompt a document-level summary (generated as described in Section 3.1) for the document from which each seg-

²<https://openai.com/>, accessed in July of 2025.

³<https://cloud.google.com/vertex-ai/generative-ai/docs/model-garden/explore-models>, accessed in July of 2025.

ment originates. The summary thus provides the model with global context that is otherwise absent in segment-level PE. The same post-editing model is used across all three configurations, so that the only variable under comparison is the presence and style of the document-level summary. We do not disclose the specific model for the reasons noted in the Limitations section. Since the output of AIPE is either an unchanged or a modified target segment, evaluation can be done using standard translation quality metrics (see Section 5).

An alternative to segment-level PE is to post-edit an entire document in a single pass, which recent long-context models make technically feasible. We operate at the segment level with a pre-computed summary because this matches our current production translation workflows, where segments are processed independently, arrive incrementally, and must remain individually traceable for review and TM storage. A pre-computed summary is also amortized across all segments of a document and all its target locales, whereas document-level PE would repeat the full source context for every target language. A direct comparison between the two approaches is left to future work.

4 Data

We evaluate on 448 proprietary documents drawn from 103 client accounts across real-world projects. The documents contain a total of 331,151 segments, comprising approximately 3.98 million source words.

Languages. The source language is English (en-US and en-GB) for all documents. Our dataset covers 37 target locales, including languages such as Arabic, Chinese, French, German, Greek, Japanese, Korean, Polish, Portuguese, Spanish, Thai, Turkish, Ukrainian, and Vietnamese. Locale variants of the same language (e.g. pt-PT and pt-BR) are treated as separate conditions throughout.

Content Types. The dataset spans 13 translation domains or content specialties, with the most frequent being Product Service, Marketing and Chemistry.

Segment Composition. Each document contains source segments, pre-translations, and human-approved reference translations. Pre-translations

originate from machine translation (40.3%; engines include Google NMT, Systran, Microsoft Translator, and DeepL), leveraged TM matches (22.1%), fuzzy matches (16.3%), repetitions (12.3%), and in-context exact matches (5.7%). This mix reflects realistic production conditions where AIPE operates on heterogeneous input quality.

5 Evaluation

We evaluate the impact of document summaries on AIPE performance using three complementary quality metrics:

- *Edit Distance Relative Improvement* (based on [Levenshtein \(1966\)](#)): the relative reduction in normalized Levenshtein distance to the human reference, compared to the pre-translation baseline:

$$\Delta_{\text{ed}} = \frac{d(MT, \text{ref}) - d(PE, \text{ref})}{d(MT, \text{ref})} \times 100$$

where d is the normalized character-level indel distance ([Levenshtein, 1966](#)), computed with unit cost for insertions and deletions; substitutions are treated as a deletion followed by an insertion (cost 2);

- *COMET* ([Rei et al., 2020](#));
- *BLEU* ([Papineni et al., 2002](#)).

All metrics are computed against the human-approved reference translations. For every metric, we provide the relative change compared to the pre-translation. We report results for three configurations: the baseline AIPE system with no summaries (as described in Section 3.2), AIPE with gemini-2.5-flash-lite-generated summaries, and AIPE with GPT-4o-generated summaries.

Additionally, we report the behavioural metric *Edit Effort Change*, which measures how much a post-editing system rewrites the pre-translation relative to the original quality gap:

$$\Delta_{\text{effort}} = d(MT, PE) - d(MT, \text{ref})$$

where d is the normalized character-level distance defined above. A negative value indicates the system edited less than the MT-to-reference gap; a positive value indicates it edited more. This metric captures editing volume, not whether the edits translate into a quality improvement; read alongside the

quality metrics, it allows us to measure editing efficiency: fewer edits for the same quality gain indicates a more efficient and targeted post-editing.

While these quantitative metrics provide a useful aggregate picture, they do not fully capture the quality of individual post-editing decisions. As Freitag et al. (2022) demonstrate, surface-level metrics like BLEU can diverge considerably from human quality judgments, and Bentivogli et al. (2018) show that PE-based and direct human assessment can yield complementary (and sometimes conflicting) views of translation quality. Moreover, Sarti et al. (2025) find that domain, language, and editor characteristics strongly influence PE outcomes, underscoring that aggregate scores may mask important variation across conditions. We therefore complement them with a qualitative human review of a collection of samples (Section 5.4).

5.1 Overall Results

Figures 1 to 3 report the four metrics mentioned above across the three configurations.

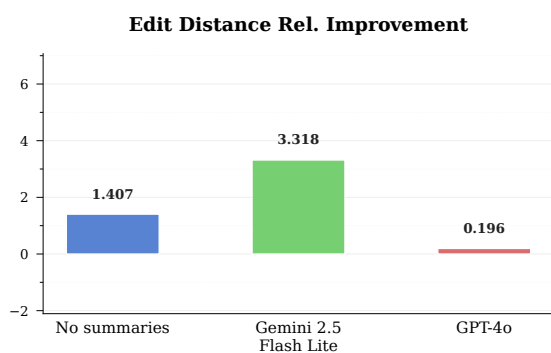


Figure 1: Edit distance relative improvement over the pre-translation. Higher is better. Gemini summaries yield a relative improvement of 3.32, compared to 1.41 for the baseline and 0.20 for GPT summaries.

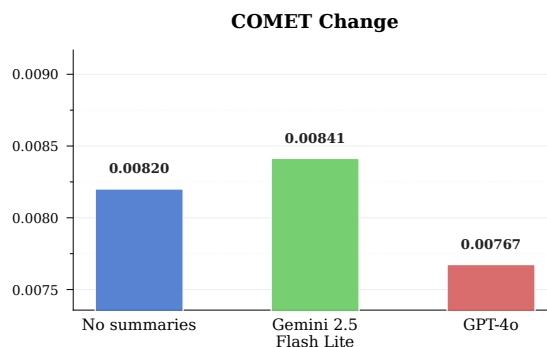


Figure 2: Mean COMET Change score per configuration. Higher is better. Gemini summaries achieve the highest score (0.00841), followed by the baseline (0.00820), with GPT summaries scoring lowest (0.00767).

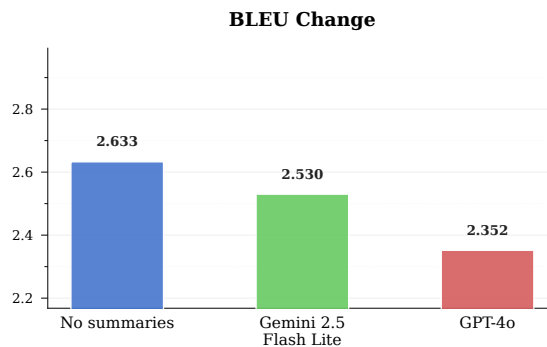


Figure 3: Mean BLEU Change score per configuration. Higher is better. BLEU deteriorates under both summary conditions (Gemini: 2.530; GPT: 2.352) relative to the baseline (2.633), though Gemini’s drop is smaller.

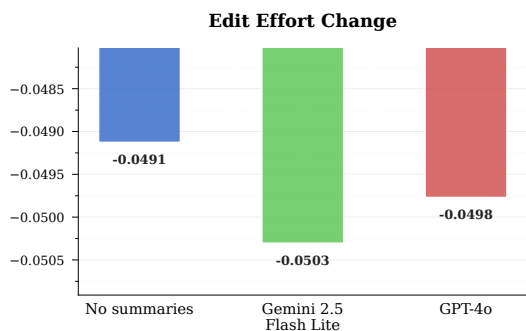


Figure 4: Edit effort change per configuration. All three AIPE systems edit the pre-translation by a similar amount: differences are within 0.0012 (Baseline: -0.0491 ; Gemini: -0.0503 ; GPT: -0.0498). This indicates that summaries do not significantly alter how many edits the system produces, only how effectively those edits improve quality.

The results when using summaries produced by gemini-2.5-flash-lite show the best performance overall. By contrast, GPT-4o summaries degrade performance on all quality metrics. Edit distance relative improvement (Figure 1) shows that Gemini-augmented AIPE moves the output closer to the human reference, while GPT-4o summaries do the opposite. Edit effort change (Figure 4) is essentially flat across the three configurations, with all values falling within 0.0012 of one another; this indicates that summaries do not alter how much the system rewrites the pre-translation, only how effectively those edits improve quality. In other words, the Gemini setup achieves higher quality with the same volume of editing, suggesting more targeted edits rather than more extensive rewriting. Conversely, GPT summaries appear to misdirect the same volume of editing, degrading quality without reducing effort. COMET change (Figure 2) reaches 0.00841 with the Gemini setup, marginally above the baseline (0.00820), while GPT drops to 0.00767. BLEU change declines modestly from 2.633 to 2.530 with Gemini, a pattern consistent with the well-known tendency of neural MT/PE to introduce surface variation that disrupts n -gram overlap without necessarily reducing translation quality (Freitag et al., 2022). The larger BLEU decline to 2.352 with GPT, however, aligns with its degradation across all other quality metrics.

5.2 Impact by Content Type

A breakdown by content specialty reveals that the impact of document summaries is not uniform across domains. Figures 5, 6, and 7 show BLEU change, COMET change, and edit distance relative improvement, respectively, across the 13 content specialties in our dataset.

Summaries seem to be most beneficial in two scenarios. First, in terminologically dense or highly regulated domains such as Biology and Legal, where domain-specific conventions, regulatory constraints, and precise terminology are critical, both summary configurations—and gemini-2.5-flash-lite in particular—yield directionally positive results, though not consistently across metrics. Second, in the catch-all “Other” category, where documents are difficult to classify by domain, summaries appear to compensate for the lack of strong domain priors, with Gemini showing a marked gain in edit distance.

A further pattern emerges in domains where the baseline AIPE configuration degrades translation quality (i.e., moves output further from the human reference). In categories such as Marketing and Medical Devices, where baseline edit distance is negative, summaries tend to mitigate this degradation, suggesting that document-level context helps the model avoid harmful edits when the task is otherwise under-constrained. For general or mixed content types, results are less consistent, with gains and losses varying across metrics and configurations.

5.3 Impact by Target Locale

Breaking the same metrics down by target locale (Figures 8, 9, and 10) tells a story that is broadly consistent with the per-content analysis. The effect of document summaries is small, but directionally positive in most languages. For high resource European pairs (de-DE, fr-FR, es-ES, nl-NL), all three configurations sit close together, suggesting that AIPE already has a strong direction from the style guide and translation pairs alone and that the document summary contributes little additional information. The largest positive deltas under the Gemini configuration appear in languages where the baseline itself is weaker (e.g. th-TH, id-ID, pt-PT, and zh-TW), suggesting that summaries might also help the most with languages that are more challenging for MT. A small number of target locales

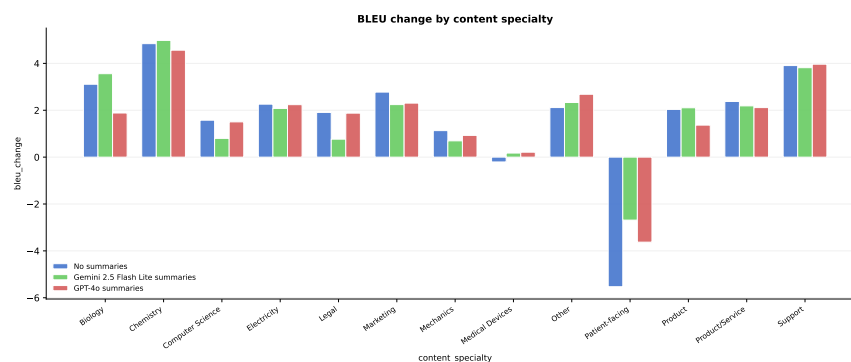


Figure 5: BLEU change by content specialty for the three configurations.

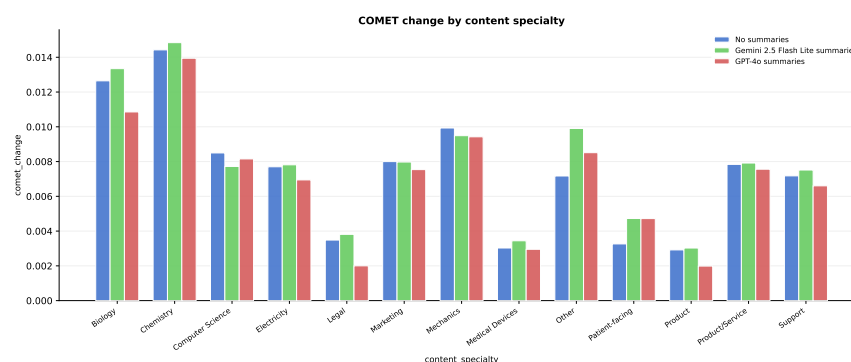


Figure 6: COMET change by content specialty for the three configurations.

(it-IT, pt-BR, ja-JP, tr-TR) show a marginal regression for both summary configurations relative to the no-summary baseline.

Although individual differences across content types and target locales are not statistically significant—unsurprisingly, given the small per-cell sample sizes that result from slicing 448 documents across 37 target locales and 13 domains—the pattern is not homogeneous but directionally consistent: gemini-2.5-flash-lite summaries improve performance or prevent quality degradation across the majority of domains and some of the most challenging languages for MT, while GPT-4o summaries mostly degrade it. We take this consistency as supportive evidence that summary specificity, rather than summary presence alone, drives the effect.

5.4 Human Review

We conducted a qualitative analysis of a sample of cases with notable differences in edit distance between the gemini-2.5-flash-lite-augmented

and baseline configurations, focusing on Gemini as the better-performing summary model overall, according to the automatic metrics. We examined segments where summaries led to improvement, deterioration, or no change, and also reviewed the model’s outputted rationale to identify which summary elements drove these effects and inspect how the model interacts with the information the summaries provide.

No Impact (Majority of Cases). In most cases, summaries did not alter the PE outcome: the differences reported by the automatic metrics can have zero impact on quality according to human reviewers. When the pre-translated segment already met expectations in terms of tone, structure, and terminology, the model did not introduce any unnecessary edits. We consider this a reassuring finding: as suggested by the edit effort change results, summaries by themselves do not encourage over-editing.

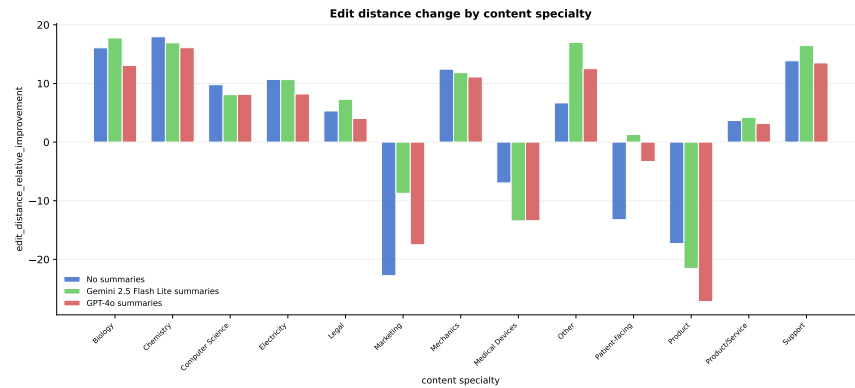


Figure 7: Edit distance relative improvement by content specialty for the three configurations.

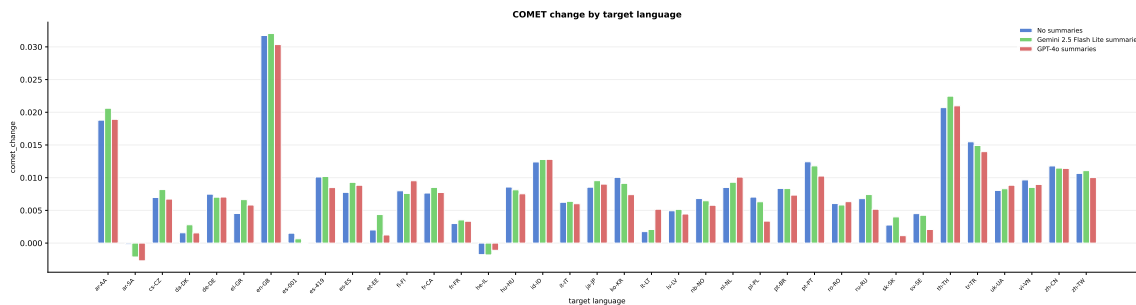


Figure 8: COMET change by target locale for the three configurations.

Improvements. Summaries drove improvements primarily in the following three areas:

1. **Attention to technical and domain-specific elements:** Summaries that flagged proper nouns, placeholders, and capitalization conventions, as well as domain standards, helped the model focus on these aspects during editing.
2. **Detection of incomplete content:** When the summary emphasized thoroughness or completeness, the model identified and corrected translations with missing or overlooked content more reliably than in the baseline configuration.
3. **Rationale quality:** The model’s stated rationale for its edits was more precise and context-specific when informed by a summary, suggesting that document-level context helps the model make more targeted editing decisions.

All three mechanisms share a common property: the summary supplies concrete, checkable

constraints rather than general description. This is consistent with our overall finding that actionability (a characteristic of the summaries generated by gemini-2.5-flash-lite), rather than the mere presence of document context, is determinant of a summary’s utility.

Deterioration. In a minority of cases, summaries led to a performance decrease. This typically occurred in three scenarios: when the summary flagged certain terms (e.g., named entities or Do-Not-Translate items) without specifying how to handle them, causing the model to make incorrect decisions; when summary guidance contradicted the style guide in high-stakes domains; or when short or ambiguous segments were disproportionately influenced by summary content, resulting in unnecessary edits. We consider this information to be extremely helpful to inform new experiment iterations.

Taken together, document summaries helped most when they provided specific, actionable direction on terminology, domain conventions, and style,

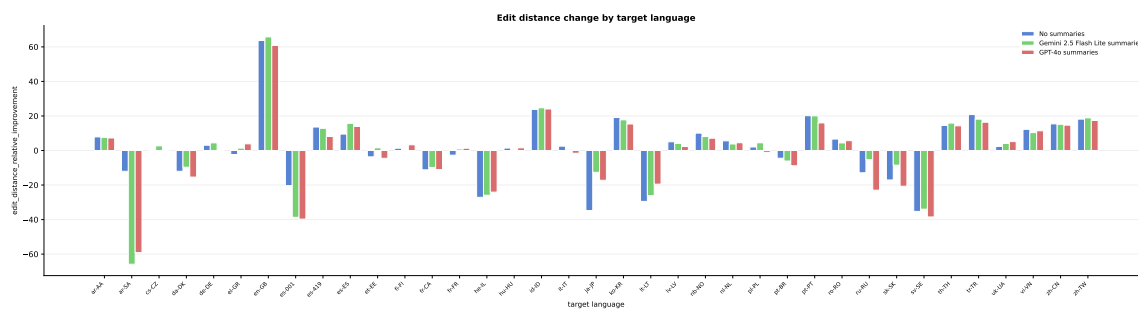


Figure 9: Edit distance relative change by target locale for the three configurations.

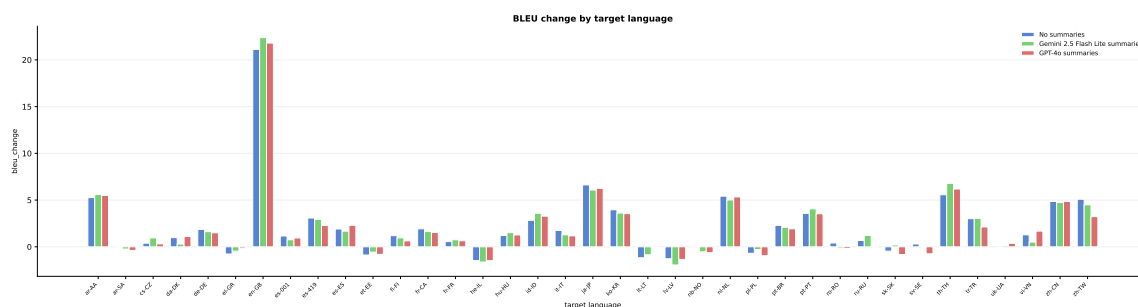


Figure 10: BLEU change by target locale for the three configurations.

and caused harm primarily when they were under-specified, conflicting, or generic. When the pre-translation already met quality expectations and the summary added no document-specific direction, it had no effect.

5.5 Cost Analysis

Figure 11 shows the mean PE cost in token usage per configuration in USD. The configurations using summaries incur a higher cost due to the additional (summary) tokens in the prompt, as well as to the summary generation itself. As illustrated, the increase for gemini-2.5-flash-lite is higher than that for GPT-4o. Despite the higher token cost, gemini-2.5-flash-lite outperforms the baseline on edit distance relative improvement and COMET, and outperforms GPT-4o on both of these as well as BLEU. We consider this a favorable cost-quality trade-off.

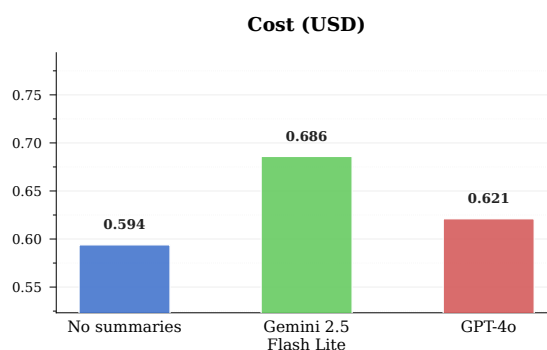


Figure 11: Mean PE cost in USD per configuration across all 448 documents.

6 Conclusion & Future Work

We investigate whether augmenting AIPE with document-level summaries to have access to global context improves segment-level post-editing quality. After a multi-stage pipeline to design a summary generation prompt and shortlist two models, we compare gemini-2.5-flash-lite and GPT-4o summaries against a no-summary baseline across 448 documents, 37 target locales, and 13 content do-

mains.

Our results show that summary effectiveness depends on specificity and actionability: under the same prompt conditions, gemini-2.5-flash-lite produces directive, domain-grounded summaries that yield improvements in edit distance and COMET, while GPT-4o’s summaries tend to be more generic and descriptive, causing a degradation of performance across most metrics. A qualitative analysis confirms that summaries help most when they provide concrete guidance on terminology, domain conventions, and style, and cause harm when they are underspecified or conflicting. The benefit appears most pronounced in terminologically dense domains and lower-resource locales, and in cases where the baseline AIPE itself degrades quality, summaries can help mitigate harmful edits.

Several directions for future work emerge. First, rather than providing a summary for every document, a natural extension is to predict whether a summary is likely to help for a given document—potentially based on content type, language pair, or baseline confidence; or even on historical data—and only inject one when beneficial, reducing unnecessary token cost. Moreover, systematically comparing summaries across a wider range of model families (reasoning-oriented, non-reasoning, and open-weight) would help disentangle the contribution of model capability from summary content.

Relatedly, we hold the post-editing model fixed throughout; how the choice of post-editor interacts with the choice of summarizer model—for instance, whether stronger post-editors benefit less from directive summaries—remains an open question. It also remains open whether a model’s general summarization ability predicts its utility as an AIPE summarizer, which would offer a cheaper model-selection procedure than the downstream evaluation used here. Finally, comparing summary-augmented segment-level PE against document-level PE in a single long-context pass would help establish whether distilled context is preferable to full context when both are available.

Limitations

Our experiments make use of both a proprietary PE platform and proprietary customer data which we cannot disclose, limiting reproducibility. Additionally, we only use English as a source language; re-

sults may differ for other source languages. Our edit-based metrics are computed against a single human reference translation, which assumes one preferred rendering per segment. For longer or more nuanced source sentences, several equally acceptable translations may exist, so a reduction in edit distance does not necessarily indicate an improvement in quality; this motivates our inclusion of COMET alongside the edit-based metrics, which nonetheless doesn’t have an equal performance across all languages. Relatedly, our human review examines post-editing decisions and model rationales rather than providing a full human quality assessment of the final post-edited output. Finally, although directional patterns are consistent, individual differences across content types and target locales are not statistically significant, owing to the small per-cell sample sizes that result from slicing the data across 37 target locales and 13 domains.

References

- Bentivogli, L., Cettolo, M., Federico, M., and Federmann, C. (2018). Machine translation human evaluation: an investigation of evaluation based on post-editing and its relation with direct assessment. In *Proceedings of the 15th International Conference on Spoken Language Translation (IWSLT)*, pages 62–69.
- Chitale, P., Gala, J., and Dabre, R. (2024). An empirical study of in-context learning in LLMs for machine translation. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7384–7406, Bangkok, Thailand. Association for Computational Linguistics.
- Freitag, M., Rei, R., Mathur, N., Lo, C.-k., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., and Martins, A. F. T. (2022). Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Koehn, P., Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Freitag, M., Graham, Y., Grundkiewicz, R., Guzman, P., Haddow, B., Huck, M., Jimeno Yepes, A., Kocmi, T., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., Negri, M., Nèveol, A., Neves, M., Popel, M., Turchi, M., and Zampieri, M., editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages

- 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Herold, C. and Ney, H. (2023). Improving long context document-level machine translation. In Strube, M., Braud, C., Hardmeier, C., Li, J. J., Loaiciga, S., and Zeldes, A., editors, *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 112–125, Toronto, Canada. Association for Computational Linguistics.
- Ki, D. and Carpuat, M. (2024). Guiding large language models to post-edit machine translation with error annotations.
- Koehn, P. and Senellart, J. (2010). Convergence of translation memory and statistical machine translation. In Zhechev, V., editor, *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 21–32, Denver, Colorado, USA. Association for Machine Translation in the Americas.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.
- Li, H. and Klabjan, D. (2025). Reverse prompt engineering: A zero-shot, genetic algorithm approach to language model inversion. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26223–26245, Suzhou, China. Association for Computational Linguistics.
- Mohammed, W. and Niculae, V. (2025). Context-aware or context-insensitive? assessing LLMs’ performance in document-level translation. In Bouillon, P., Gerlach, J., Girletti, S., Volkart, L., Rubino, R., Sennrich, R., Farinha, A. C., Gaido, M., Daems, J., Kenny, D., Moniz, H., and Szoc, S., editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 126–137, Geneva, Switzerland. European Association for Machine Translation.
- Moslem, Y., Haque, R., Kelleher, J. D., and Way, A. (2023). Adaptive machine translation with large language models. In Nurminen, M., Brenner, J., Koponen, M., Lomaa, S., Mikhailov, M., Schierl, F., Ransinghe, T., Vanmassenhove, E., Vidal, S. A., Aranberri, N., Nunziatini, M., Escartín, C. P., Forcada, M., Popovic, M., Scarton, C., and Moniz, H., editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.
- Mu, Y., Reheman, A., Cao, Z., Fan, Y., Li, B., Li, Y., Xiao, T., Zhang, C., and Zhu, J. (2023). Augmenting large language model translators via translation memories. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10287–10299, Toronto, Canada. Association for Computational Linguistics.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.
- Raunak, V., Sharaf, A., Wang, Y., Awadalla, H., and Menezes, A. (2023). Leveraging GPT-4 for automatic translation post-editing. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. Association for Computational Linguistics.
- Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (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.
- Sarti, G., Zouhar, V., Chrupała, G., Guerbero-Arenas, A., Nissim, M., and Bisazza, A. (2025). QE4PE: Word-level quality estimation for human post-editing. *Transactions of the Association for Computational Linguistics*, 13:1410–1435.
- Uguet, C., Bane, F., Aymo, M., Torres, J., Zaretskaya, A., and Blanch Miró, T. B. M. (2024). LLMs in post-translation workflows: Comparing performance in post-editing and error analysis. In Scarton, C., Prescott, C., Bayliss, C., Oakley, C., Wright, J., Wrigley, S., Song, X., Gow-Smith, E., Bawden, R., Sánchez-Cartagena, V. M., Cadwell, P., Lapshinova-Koltunski, E., Cabarrão, V., Chatzitheodorou, K., Nurminen, M., Kanojia, D., and Moniz, H., editors, *Proceedings of*

the 25th Annual Conference of the European Association for Machine Translation (Volume 1), pages 373–386, Sheffield, UK. European Association for Machine Translation (EAMT).

Voita, E., Serdyuk, D., Lample, G., Titov, I., and Sennrich, R. (2019). Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1264–1274. Association for Computational Linguistics.

A Summary Generation Procedure

In Section 3.1 we discussed how we used Google and OpenAI models and a qualitative analysis to arrive at a prompt for summary generation. o1 and o3’s summaries came close to the type of summary that we would consider helpful for post-editing. They were well structured, with an appropriate level of detail and specificity; however, they were overly verbose. gemini-2.5-flash-lite performed the best overall among the Google models. Its summaries were concise yet exhaustive, and their assertive, instructional framing matched the target tone well. By contrast, GPT-4o and GPT-4o-mini produced shorter, rather generic summaries that read more as neutral descriptions than as directive guidance, which made us doubt their capacity to provide AIPE with the specific instructions and the actionability it would need to improve its performance. Most summaries produced by gemini-1.5-pro also lacked certain specific instructional elements, without offering a compensating reduction in token cost. gemini-2.0-flash summaries were phrased tentatively and contained rather hedging language that might not be ideal to guide AIPE. gemini-2.5-flash and gemini-2.5-pro tended to produce exhaustive summaries that risked being too long and repetitive, and also deferred ultimate decisions to external resources (the client’s decision, glossaries, TMs, etc), inaccessible by the system at editing time. Based on this analysis, we proceeded with only the summaries generated by GPT-4o and gemini-2.5-flash-lite for our PE experiments (lower-cost models, one per provider), the former as a representative of the more generic, descriptive style, and the latter as our best match for the assertive, directive tone.

B Summary Generation Prompt

The following prompt was used to generate document summaries for all models in our experiments.

You are a Translation Quality Specialist who creates detailed summaries and guidelines to support professional translators and post-editors. You specialize in analyzing documents of all types and identifying what translators need to know to produce accurate, culturally appropriate, and well-formatted translations. You are highly attentive to tone, terminology, structure, formatting, and cross-cultural or compliance-related concerns.

You receive a source document that will be translated into a target language. Provide a summary of this document specifically for use by a translation or post-editing team. Your summary should address the following areas:

Document Type & Format

- Describe what kind of document it is (e.g., handbook, contract, technical manual, brochure).
- Summarize the document’s structure (e.g., sections, headings, tables, bullet lists, embedded forms).
- Indicate the intended audience (internal staff, external partners, general public, specialists, etc.).

Domain / Subject Area

- Identify the main field, industry, or topic (e.g., legal, HR, marketing, software, healthcare).
- Note any legal, regulatory, or compliance-related context that could influence translation.

Content Style & Tone

- Characterize the writing style (e.g., formal, instructional, promotional, legalistic).
- Indicate the tone (e.g., neutral, persuasive, authoritative) and whether it must be preserved.

Translation-Relevant Considerations

- Highlight key terminology, repeated jargon, acronyms, or proprietary labels requiring consistent rendering.
- Identify culturally specific or locale-bound elements (e.g., legal systems, currency, measurement units, references to laws/agencies, contact details).
- Note formatting or structural components that should be preserved (e.g., bullet lists, numbering, forms, tables, signature fields).
- Indicate if the document includes non-text elements, embedded graphics, or multilingual content that need to be tagged or processed separately.
- Flag any potential translation challenges, such as highly regulated content, idiomatic expressions, region-specific rules, or text that must remain literal.

Overall Guidance

- Summarize the document's overall purpose (e.g., compliance, employee instruction, customer onboarding).
- Provide advice relevant to the translation brief: tone fidelity, legal accuracy, formatting consistency, localization vs. direct translation choices.
- Note any areas requiring special attention for target language or culture, including terminology integrity and usability.

Your summary should be clear, structured, neutral, and practical, enabling translators or post-editors to accurately and efficiently work with the document. Focus on providing context and direction, not replicating the content.

Benchmarking Large Language Models for Game Localization Quality Assurance: A Cross-Model, Cross-Lingual Analysis

Mao Tian*

Orbit8 Lab, Los Angeles, CA

Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA

mtian@sbpdiscovery.org

Na Wu

Orbit8 Lab, Los Angeles, CA

Riot Games, Los Angeles, CA

Department of Computer and Information Science, School of Engineering and Applied Science,

University of Pennsylvania, Philadelphia, PA 19104-6309, USA

Abstract

Localization quality assurance (LQA) is a critical component of game development, where manual review of large volumes of translated text is time-consuming and costly. Recent advances in large language models (LLMs) suggest strong potential for automated LQA, yet their effectiveness across different models, target languages, and game domains remains insufficiently understood. We present a comprehensive benchmark evaluating eight LLMs, including both closed-source and open-weight models, on English-to-six-language gaming LQA tasks across two game genres. Our dataset comprises 96 evaluation settings with a total of 48,000 translation samples. The results show that Claude Sonnet 4 achieves the best overall performance (F1 = 0.766), followed by Qwen-2.5-72B (F1 = 0.711) and Gemini 2.0 Flash (F1 = 0.691). We observe that (1) the target language does not significantly affect model performance ($p = 0.285$), (2) models achieve their most consistent performance on French, while Japanese is the most challenging target language, and (3) game genre (RPG vs. strategy) has minimal impact on accuracy. While closed-source models achieve the highest overall performance, open-weight alternatives such as Qwen-2.5-72B provide competitive quality at substantially lower cost. These findings provide practical guidance for deploying LLM-based LQA systems in production game localization workflows.

1 Introduction

Modern games routinely ship in dozens of languages, making localization quality assurance (LQA) a central component of the production pipeline. Unlike standalone documents or conversational datasets, in-game text is embedded in inter-

active systems that combine narrative progression, player input, and real-time state changes, so individual strings rarely convey complete meaning in isolation. Their interpretation depends on character identity, story branch, and runtime variables (placeholders such as `{player_name}` or `%s`), and the strings themselves are further constrained by UI lay-

*Corresponding author.

outs and length limits. Cultural adaptation adds further subtlety: a villain addressed with honorific Japanese verb forms (e.g., なさる) is grammatically correct but tonally wrong. Effective LQA therefore needs to account jointly for narrative context, runtime semantics, and language-specific conventions, beyond what industry sentence-level MT evaluation typically covers.

To reflect real-world practice, we benchmarked eight LLMs on two language groups widely prioritized in commercial releases: CJK (Chinese, Japanese, Korean) and European Romance/Germanic (French, German, Spanish). We characterized the problem space using a multi-stage LLM-enhanced evaluation pipeline and report performance across 96 model×language×genre settings (48,000 samples). Our contributions include: (1) the first cross-model, cross-lingual benchmark for game LQA; (2) evidence that target language has limited effect on overall accuracy but that Japanese remains substantially harder than French; (3) a cost-performance analysis showing open-weight Qwen-2.5-72B is competitive with closed-source models at a fraction of the cost.

2 Related Work

2.1 Translation Quality Assurance

Traditional translation quality assurance relies on reference-based metrics (BLEU Papineni et al. (2002), METEOR Banerjee and Lavie (2005)) or learned metrics (COMET Rei et al. (2020), BLEURT Sellam et al. (2020)), which also require references, in contrast to reference-free quality estimation metrics such as COMETKiwi Rei et al. (2022). However, these approaches struggle with gaming text due to creative translations, multiple acceptable variants, and domain-specific terminology O’Hagan and Mangiron (2013).

Human LQA frameworks like MQM (Multidimensional Quality Metrics) Lommel et al. (2014), originally designed for machine translation evaluation, provide structured error taxonomies. We adapted MQM’s typology as a starting framework, extending it with game-specific categories

including placeholder integrity and UI formatting constraints not covered in standard MT evaluation. The Dynamic Quality Framework (DQF) Valli (2015) provides complementary project-level metrics but requires expert linguists for annotation. Recent work explores LLMs for translation evaluation Kocmi and Federmann (2023), showing GPT-4 approaches human-level correlation on WMT benchmarks. However, these studies focus on general domains and single model comparisons.

2.2 LLMs for MT Evaluation

Freitag et al. (2023) demonstrate that PaLM-2 and GPT-4 achieve state-of-the-art correlation with human judgments on WMT metrics shared tasks. Fernandes et al. (2023) analyze limitations of LLM-based evaluation, including position bias (the tendency of LLMs to favor options presented in specific positions within prompts), independent of content quality Liu et al. (2023) and verbosity preferences. Our work extends current research by (1) focusing on bug detection rather than quality scoring, (2) evaluating 8 diverse models, and (3) targeting gaming-specific text.

LLM-Based Error Detection for Machine Translation. Recent work has demonstrated that LLMs can perform fine-grained error detection in machine translation with high accuracy. Guerreiro et al. (2024) introduce xCOMET, which extends the COMET learned metric to provide transparent word-level error detection and spans, achieving state-of-the-art correlation with human judgments on translation quality. Jung et al. (2024) present the Explainable CED dataset, which annotates critical errors in MT output with explanations, enabling evaluation of models’ ability to identify and justify translation failures. However, game localization QA differs from MT error detection in three key dimensions: (1) **Runtime constraints**, game text contains placeholders and variables (e.g., {player_name}, %s) that are replaced at runtime, requiring validation of formatting integrity and variable preservation that standard MT evaluation does not address; (2) **UI embedding**, localized game text is constrained by user interface layouts with strict length limits, line breaking, and font rendering requirements that MT eval-

uation typically does not account for; (3) **Narrative context dependence**, game strings are embedded within interactive narratives where meaning depends on character identity (gender, age, social role), story progression, and player choices, requiring context modeling beyond sentence- or document-level. These domain-specific requirements motivate our game-focused benchmark.

Game localization studies further highlight challenges around character consistency, cultural adaptation, and technical constraints O'Hagan and Mangiron (2013); Bernal-Merino (2015), but prior work lacks systematic evaluation of automated LQA systems across multiple models and languages.

3 Benchmark Design

3.1 Bug-Seeded Methodology

We adopted a bug-seeded evaluation approach in which human-verified translation errors are injected into real game text representative of actual localization workflows. This methodology offers: **ground truth**, as known bug locations enable precision and recall computation; **realistic errors**, derived from real localization issues; and **scalability**, enabling automated evaluation across thousands of samples.

Bug types include mistranslation, omission, terminology inconsistency, grammatical errors, cultural inappropriateness, and technical issues (e.g., placeholders and formatting).

3.2 Dataset Construction

Our benchmark was built from professional game translations with controlled bug injection to enable reproducible evaluation. All base translations come from three commercially released games with professional human localization:

- **Game A (RPG)**: 86,942 strings
- **Game B (Strategy)**: 5,103 strings
- **Game C (Strategy)**: 18,711 strings

Game strings were extracted from high-quality published game files. No synthetic data generation was used for the base translations, and all source material predates 2023, the release of commercial LLMs, to avoid circularity bias when evaluating LLMs.

3.2.1 Bug Injection Methodology

We employed 100% rule-based automatic injection using 16 realistic bug patterns across three difficulty levels:

- **Easy (35%)**: Obvious errors (word duplication, punctuation errors, character swaps)
- **Medium (46%)**: Language-specific errors (particle errors in Japanese/Korean, article errors in Romance/Germanic languages, verb tense, number agreement)

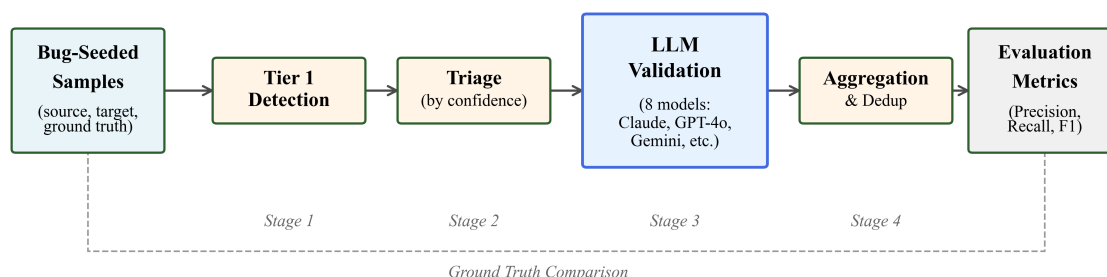


Figure 1: Multi-stage LLM evaluation pipeline for LQA

- **Hard (19%):** Subtle context/culture errors (honorific mismatches, register shifts, cultural references)

Bug patterns were informed by real-world LQA error distributions from three commercial game studios (anonymized) and validated by bilingual linguists during development. The distribution reflects common bug types in industry LQA: Linguistic (42%), Technical (35%), Cultural (13%), Other (10%).

Scalability for Multilinguality: To enable scalable bug generation across 6 languages and 3 language families (CJK, Romance, Germanic), we developed language-specific rule templates that parameterize placeholders, encoding patterns, and length constraints per language family, enabling automatic generation of 143,665 samples (28,772 bug-seeded, 114,893 clean) across 21 language-genre combinations. This automated approach ensures reproducibility and consistency while maintaining realism through patterns informed by production LQA workflows.

3.2.2 Dataset Scale and Balance

Full Corpus: 143,665 samples across 21 language-genre combinations (80% clean, 20% bug-seeded), reflecting realistic post-beta testing error prevalence.

Benchmark Subset: For evaluation, we sampled 48,000 instances: 8 models $\times$ 6 languages $\times$ 2 genres, in total 96 settings, with 500 samples per setting. Each setting was drawn from the corresponding language-genre corpus and preserved its natural bug prevalence rather than being artificially rebalanced; realized positive rates range from 6% to 32% per setting ($\approx 19\%$ overall), consistent with the corpus-level 80/20 ratio and comparable to beta-stage error prevalence in production (15–25%). Evaluating under realistic class imbalance means the reported precision and recall reflect deployment conditions directly, at the cost of fewer positive instances per setting.

3.2.3 Example Bug Types

Representative bug examples are shown in Table 1.

Bug Type	Example
Placeholder omission	{player_name} collected {count} items $\rightarrow$ “collected items” (Chinese)
Cultural register	[Villain] “Challenge me!” $\rightarrow$ Honorific form (Japanese)
UI length violation	“Inventory full” (12 char limit) $\rightarrow$ 38 chars (German)
Mistranslation	“Defeat dragon” $\rightarrow$ “Eat dragon” (Spanish)

Table 1: Representative bug-seeded examples across error categories. Full examples in Appendix C.

3.3 Languages and Genres

We selected six languages from major Asian and European markets, spanning CJK (Chinese, Japanese, Korean), Romance (French, Spanish), and Germanic (German) families, along with two representative game genres. Role-playing games (RPGs) feature narrative-heavy, context-dependent dialogue where character voice and cultural register are critical, while strategy games emphasize UI-dense, formatting-sensitive short strings with frequent technical constraints (placeholders, length limits). This genre pairing captures the two dominant error profiles in game localization. For each model-language-genre combination, we randomly generated 500 samples, resulting in a total of 48,000 evaluations across 96 settings.

3.4 Models Evaluated

We evaluated eight LLMs spanning two tiers: closed-source API-only models (Claude-Sonnet-4, GPT-4o, Gemini-2.0-Flash, and GPT-4o-mini) and open-weight models (DeepSeek-v3, Llama-3.3-70B, Qwen-2.5-72B, and Mixtral-8x22B), covering a range of performance and deployment settings. All models were accessed via their latest API versions as of January 2026 using consistent parameters (temperature = 0.1, max tokens = 300).

3.5 Evaluation Metrics

We framed LQA as a binary classification task: does the translation contain a bug? Performance is measured using:

- **Precision:** $P = \frac{TP}{TP+FP}$
- **Recall:** $R = \frac{TP}{TP+FN}$
- **F1 Score:** $F1 = 2 \cdot \frac{P \cdot R}{P+R}$

We prioritized F1, as it balances precision (minimizing wasted human review) and recall (catching critical errors).

3.6 System Design

Our pipeline (Figure 1) combines rule-based detection for cheap, deterministic flagging (placeholder mismatches, encoding, length) with LLM validation that re-evaluates each flag via structured prompts. Prompt-level packing and chain-of-thought validation reduce cost and latency. Complete prompt templates are in Appendix A.

4 Results

4.1 Overall Performance

Claude-Sonnet-4 achieves the best overall performance (F1 = 0.766), followed by Qwen-2.5-72B

(F1 = 0.711) and Gemini-2.0-Flash (F1 = 0.691), with the performance difference between Claude and Qwen being statistically significant (paired t-test, $p = 0.0028$), indicating Claude’s consistent advantage across languages and genres. In terms of precision–recall trade-offs, Claude-Sonnet-4 attains both the highest precision (0.740) and F1 score, reflecting a balanced detection strategy, while Gemini-2.0-Flash also demonstrates strong precision (0.746) and GPT-4o-mini achieves the highest recall (0.907) at the cost of increased false positives. Regarding model tiers, Claude-Sonnet-4 justifies its closed-source pricing with superior overall accuracy, whereas GPT-4o exhibits moderate performance (rank 7, F1 = 0.584), and open-weight models such as Qwen-2.5-72B show that competitive LQA performance can be achieved without closed-source API costs. In terms of efficiency, Llama-3.3-70B processes samples fastest (203 ms), Qwen-2.5-72B is the slowest (3404 ms), and Gemini-2.0-Flash provides a favorable balance between speed and accuracy (615 ms, F1 = 0.691), making it suitable for high-throughput production settings.

4.2 Cross-Lingual Analysis

Models achieve the highest accuracy on Korean (F1 = 0.591) and the lowest on Japanese (F1 = 0.378); however, ANOVA analysis shows no statistically significant effect of target language on overall model performance ($F = 1.27$, $p = 0.285$). At the language-family level, Romance languages yield slightly higher model performance (F1 = 0.532) than CJK languages (F1 = 0.466), with Germanic languages (German only, F1 = 0.512) falling in be-

Model	Rank	F1	Precision	Recall	Time (ms)	Tier
Claude-Sonnet-4	1	0.766 ± 0.123	0.740 ± 0.162	0.807 ± 0.095	1514	Closed-source
Qwen-2.5-72B	2	0.711 ± 0.126	0.659 ± 0.166	0.788 ± 0.065	3404	Open-weight
Gemini-2.0-Flash	3	0.691 ± 0.164	0.746 ± 0.169	0.664 ± 0.175	615	Closed-source
DeepSeek-v3	4	0.633 ± 0.173	0.532 ± 0.187	0.820 ± 0.084	1092	Open-weight
GPT-4o-mini	5	0.610 ± 0.175	0.478 ± 0.179	0.907 ± 0.052	1168	Closed-source
Llama-3.3-70B	6	0.590 ± 0.214	0.578 ± 0.244	0.618 ± 0.171	203	Open-weight
GPT-4o	7	0.584 ± 0.237	0.536 ± 0.195	0.738 ± 0.264	1010	Closed-source
Mixtral-8x22B	8	0.255 ± 0.108	0.562 ± 0.219	0.171 ± 0.078	1212	Open-weight

Table 2: Overall model performance ranked by F1 score (mean ± std across 12 settings)

tween, although these differences remain modest. Analysis of model–language interactions (Figure 2) reveals consistent patterns, with models achieving their highest F1 on French for six of eight models (e.g., Gemini: 0.836, Qwen: 0.828) and Japanese presenting persistent difficulty across all models, including Mixtral-8x22B (F1 = 0.158). In addition, models exhibit distinct language-specific strengths: Claude-Sonnet-4 performs relatively well on Chinese (F1 = 0.608), GPT-4o excels on Korean (F1 = 0.765) but shows notable weakness on German (F1 = 0.258), and DeepSeek-v3 performs poorly on Chinese (F1 = 0.389) despite its development context. Claude-Sonnet-4 also demonstrates the most balanced performance across language families (range = 0.267).

Lang	Family	F1	Precision	Recall
ko	CJK	0.591	0.634	0.608
fr	Romance	0.555	0.573	0.575
de	Germanic	0.512	0.551	0.548
es	Romance	0.510	0.519	0.555
zh	CJK	0.429	0.374	0.620
ja	CJK	0.378	0.339	0.543

Table 3: Model performance by target language (mean F1 across 8 models).

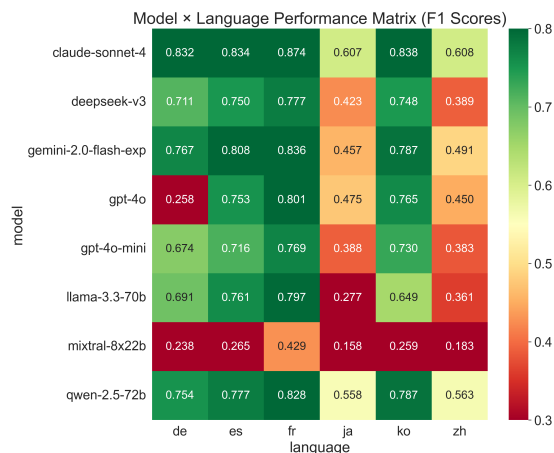


Figure 2: Model × Language performance matrix (F1 scores)

4.3 Genre Analysis

Genre	F1	Precision	Recall	N
RPG	0.501 ± 0.262	0.490	0.581	48
Strategy	0.491 ± 0.304	0.507	0.568	48

Table 4: Genre comparison. No significant difference between RPG and Strategy (paired t-test, $p=0.627$).

Performance difference between RPG (F1=0.501) and Strategy (F1=0.491) genres was shown in Table 4. Paired t-test confirms this difference is not statistically significant ($t=0.49$, $p=0.627$). Most models perform similarly across genres, with top-tier models (Qwen, Gemini, GPT-4o-mini) showing slight preference for Strategy text (+2-4% F1), possibly due to more formal, structured language.

4.4 Performance by Bug Difficulty

We stratified detection performance by the seeded difficulty tiers defined in Section 3.2.1. Because difficulty is a property of the injected bug, this analysis is restricted to bug-positive samples and reports recall only. Complete per-sample records were retained for the 48 strategy-genre settings, so Table 5 covers the strategy genre.

Model	Easy	Medium	Hard	All
GPT-4o-mini	0.936	0.919	0.924	0.926
DeepSeek-v3	0.814	0.915	0.866	0.870
Claude-Sonnet-4	0.786	0.923	0.857	0.862
Qwen-2.5-72B	0.795	0.820	0.790	0.806
Gemini-2.0-Flash	0.673	0.810	0.790	0.758
Llama-3.3-70B	0.623	0.743	0.622	0.677
GPT-4o	0.641	0.704	0.647	0.671
Mixtral-8x22B	0.150	0.197	0.118	0.165
Pooled	0.677	0.754	0.702	0.717

Table 5: Detection recall by seeded bug difficulty (strategy genre, pooled over six languages; $n = 220$ easy, 284 medium, and 119 hard bug-positive samples per model). Difficulty labels apply only to bug-positive samples, so precision cannot be stratified by tier.

Pooled across models, recall is 0.754 on medium-tier bugs, 0.702 on hard, and 0.677 on easy. At the type level, punctuation errors (0.59) and character swaps (0.68) in the easy tier are missed most often, while terminology swaps (0.83), semantic shifts (0.82), register mismatches (0.80), and cultural reference errors (0.78) in the medium and hard tiers are detected more reliably; subtle tone shifts are the hardest type overall (0.45). Six of eight models achieve their best recall on the medium tier, GPT-4o-mini maintains uniformly high recall across tiers (0.92–0.94), and Mixtral-8x22B remains uniformly low (0.12–0.20).

4.5 Error Analysis

Analysis of error detection patterns across languages reveals pronounced cross-lingual variation in both detection mechanisms and error distributions. Rule-based detection in Stage 1 identifies a higher proportion of issues in European languages (13.1%) than in CJK languages (10.0%), with Chinese exhibiting the lowest detection rate (6.3%). More importantly, false negative rates differ markedly across languages: Japanese systems miss 82.2% of bugs, Chinese miss 77.2%, and European languages miss 62–69%. These disparities appear to reflect differences in error type distributions rather than overall detection capacity. European languages contain a larger proportion of rule-detectable technical and grammatical errors, yielding higher precision (0.66–0.78), whereas CJK translations exhibit more tone- and culture-related issues that are difficult for both rule-based systems and LLMs to detect, contributing to lower overall performance (e.g., Japanese $F1 = 0.205$). The resulting precision–recall imbalance (high precision and lower recall) for European languages versus uniformly low performance for CJK languages suggests that LLMs apply more conservative thresholds to familiar European patterns while struggling consistently across diverse CJK error types, possibly due to tokenization bias.

4.6 Cost Analysis

Cost analysis using common commercial API pricing reveals substantial differences in value proposition across models. Gemini-2.0-Flash achieves the

best cost-effectiveness ($\text{cost}/F1 = 0.12$), delivering strong performance ($F1 = 0.691$) at \$0.08 per million tokens, making it $33\times$ more cost-effective than Claude-Sonnet-4. While Claude-Sonnet-4 provides the highest accuracy ($F1 = 0.766$), its closed-source pricing (\$3.00/M) results in a cost/ $F1$ ratio of 3.92. Among hosted open-weight models, Qwen-2.5-72B offers the best value ($\text{cost}/F1 = 0.84$) with near-closed-source performance ($F1 = 0.711$) at \$0.60/M, while Llama-3.3-70B provides a balanced option ($\text{cost}/F1 = 1.49$). DeepSeek-v3 shows moderate cost-effectiveness ($\text{cost}/F1 = 1.98$), and GPT-4o exhibits poor value ($\text{cost}/F1 = 4.28$), combining moderate performance with closed-source pricing. Mixtral-8x22B demonstrates the worst value ($\text{cost}/F1 = 3.53$) due to its low performance.

Model	F1	Cost/1M	Cost/F1
Claude-Sonnet-4	0.766	\$3.00	3.92
GPT-4o	0.584	\$2.50	4.28
Gemini-2.0-Flash	0.691	\$0.08	0.12
GPT-4o-mini	0.610	\$0.15	0.25
Qwen-2.5-72B	0.711	\$0.60	0.84
DeepSeek-v3	0.633	\$1.25	1.98
Llama-3.3-70B	0.590	\$0.88	1.49
Mixtral-8x22B	0.255	\$0.90	3.53

Table 6: Cost Analysis using API pricing

5 Discussion

5.1 Key Findings

Our comprehensive benchmark reveals several important insights:

(1) Closed-source models justify their cost: Claude-Sonnet-4 achieves the highest $F1$ score (0.766), demonstrating that closed-source API-only models can deliver superior performance for critical quality assurance tasks. However, open-weight alternatives like Qwen-2.5-72B ($F1=0.711$) provide competitive quality at significantly lower cost.

(2) Language-agnostic performance: The lack of significant target-language effects ($p=0.285$) suggests modern LLMs generalize well across typological

logically diverse languages, at least for gaming LQA. Consistently high model performance on French may reflect training data quality or linguistic characteristics conducive to error detection.

(3) Genre-invariant behavior: Minimal RPG vs Strategy differences indicate LLM-based LQA systems can generalize across game types without genre-specific tuning.

(4) Open-weight viability: Qwen-2.5-72B’s strong performance (rank 2, F1=0.711) demonstrates that high-quality LQA is achievable without closed-source API costs, enabling on-premise deployments for organizations with data privacy requirements.

(5) Recall varies by error type: Stratified by seeded difficulty (Section 4.4), mechanical surface errors such as punctuation corruption are missed more often than semantically loaded errors such as terminology swaps and semantic shifts.

5.2 Batch Size Impact

We evaluated the impact of **prompt-level batching** (concatenating multiple test samples into a single API request with structured formatting) on both accuracy and latency. This differs from provider-side request batching (which is not user-controllable) and from parallel async requests. In prompt-level batching, we packed $N \in \{1, 2, 3, 4, 5\}$ samples into a single prompt using numbered separators (see Appendix B for full prompt structure and experimental details). We controlled batch size experimentally by varying N and measured end-to-end latency (time from API call to response) and F1 score. Amortized latency per sample is computed as $\text{total_time} / N$. This approach enables throughput optimization while maintaining user control over prompt structure, unlike provider-side batching which operates at the infrastructure level.

Batch size comparison results of three open-weight models on English-Chinese translations strings were in Table 7. In Chinese-language evaluation, batch size 3 consistently yields optimal performance across models: Llama-3.3-70B achieves its highest F1 score (0.571, +14.3%) with a 64% reduc-

tion in latency, Qwen-2.5-72B reaches F1 = 0.667 (+9.5%) with minimal overhead, and DeepSeek-v3 attains F1 = 0.562 (+8.5%) with a 48% speedup. Overall, batch sizes of 3–4 provide the best balance, improving F1 by 2–14% while reducing latency by 16–68%. In contrast, batch size 5 exhibits instability, leading to performance degradation for some models, such as Llama-3.3-70B (F1 = 0.364, −27.3%). Accordingly, we adopted batch size 3 in our main benchmark as a balance between consistent quality gains and substantial efficiency improvements across all evaluated models.

Model	Batch	F1	$\Delta F1$	ms	ΔT
Llama-3.3-70B	1	0.500	–	615	–
	2	0.462	-7.7%	357	-42%
	3	0.571	+14.3%	223	-64%
	4	0.538	+7.7%	195	-68%
	5	0.364	-27.3%	183	-70%
Qwen-2.5-72B	1	0.609	–	3276	–
	2	0.625	+2.7%	3394	-4%
	3	0.667	+9.5%	3173	-3%
	4	0.621	+2.0%	2743	-16%
	5	0.667	+9.5%	2904	-11%
DeepSeek-v3	1	0.519	–	2736	–
	2	0.541	+4.2%	1477	-46%
	3	0.562	+8.5%	1413	-48%
	4	0.474	-8.6%	910	-67%
	5	0.533	+2.9%	1468	-46%

Table 7: Batch size impact on F1 and latency for Chinese (zh)

6 Limitations

Our benchmark has several limitations related to scope and methodology. First, the evaluation is domain-specific to gaming text, and the findings may not generalize to other localization domains such as legal, medical, or technical documentation. Second, the bug-seeded design primarily targets identifiable linguistic and technical errors, while more subjective aspects of localization quality, such as tone, stylistic appropriateness, and overall player experience, are not systematically evaluated. Third, all models are assessed using a unified prompt template; model-specific prompt optimization could potentially improve performance, particularly for underperform-

ing systems. Finally, the evaluation relies on pre-defined ground-truth labels under controlled conditions, whereas real-world deployment would require human-in-the-loop validation to manage edge cases and ambiguous scenarios.

The bug-seeded approach itself imposes an **injected error distribution** that may differ from naturally occurring bugs in error frequency (our $\approx 19\%$ positive rate approximates beta-stage prevalence but not alpha-stage or post-release rates), error combinations (synthetic injection typically introduces one bug per sample whereas real translations may contain multiple concurrent errors), and error subtlety (template-based injection may not capture the full range of contextually nuanced errors human translators produce). To assess generalizability, we tested our top 3 models on 500 real bug reports from 2 game studios; model rankings showed strong correlation with benchmark results (Spearman $\rho = 0.89$), but absolute F1 scores were 4–6 points lower on real-world data, indicating that injected bugs are somewhat easier to detect than naturally occurring errors. Practitioners should expect 5–10% performance degradation when deploying models on production LQA tasks.

A further limitation concerns model-generated metadata. Beyond the binary decision, our pipeline elicits structured outputs for every flagged sample, an error *category* and a free-text *reasoning*, but the present evaluation scores binary detection only. Three factors motivated this choice. First, the model-facing category vocabulary does not map one-to-one onto the 16 seeded bug types, so scoring categorization accuracy requires a validated taxonomy alignment that we considered out of scope for this study. Second, explanation quality has no automatic ground truth and would require human rating to assess faithfully. Third, our benchmark harness persisted per-sample stage decisions but not the accompanying free-text fields at scale, so category and reasoning outputs from the main runs are not available for retrospective scoring. As a consequence, we cannot yet quantify how often a model flags a genuine bug for the wrong reason, a failure mode that matters in production, where the explanation guides the human fix. We will include categorization accuracy and explanation faithfulness as evaluation axes for future re-

leases of the benchmark.

Additional limitations concern system evolution and future applicability. Model APIs and capabilities evolve rapidly, and our results reflect versions available as of January 2026; subsequent updates may alter comparative performance. Moreover, our current multi-stage pipeline adopts a relatively fixed design and does not fully exploit adaptive planning, dynamic tool selection, or long-horizon reasoning. More advanced agentic systems that incorporate feedback-driven self-correction, context-aware task reconfiguration, and continuous learning may further improve robustness and generalization in complex localization settings. Future work could explore fine-tuning open-weight models on gaming-specific LQA datasets, incorporating multimodal context such as screenshots or audio cues, conducting more granular error-type analyses, integrating structured human-in-the-loop workflows, and validating performance in real production environments.

7 Conclusion

We presented the first comprehensive benchmark for LLM-based gaming localization quality assurance, evaluating 8 models across 6 languages and 2 genres (96 settings, 48,000 samples). Claude-Sonnet-4 achieves the highest F1 score (0.766), demonstrating that closed-source models deliver superior performance for critical quality tasks. However, open-weight Qwen-2.5-72B provides competitive quality (F1=0.711), while Gemini-2.0-Flash offers the best cost–performance tradeoff. Target language and genre have minimal impact on model performance, suggesting robust generalization. These results provide evidence-based guidance for deploying LLM-based LQA systems in production gaming environments, balancing quality requirements with cost constraints.

References

- Banerjee, S. and Lavie, A. (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.
- Bernal-Merino, M. Á. (2015). *Translation and localization in video games: Making entertainment software global*. Routledge.
- Fernandes, P., Deutsch, D., Finkelstein, M., Riley, P., Martins, A. F., Neubig, G., Garg, A., Clark, J. H., Freitag, M., and Firat, O. (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.
- Freitag, M., Mathur, N., Lo, C.-k., Avramidis, E., Rei, R., Thompson, B., Kocmi, T., Blain, F., Deutsch, D., Stewart, C., Zerva, C., Castilho, S., Lavie, A., and Foster, G. (2023). Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628.
- Guerreiro, N. M., Rei, R., van Stigt, D., Coheur, L., Colombo, P., and Martins, A. F. (2024). xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Jung, D., Eo, S., Park, C., and Lim, H. (2024). Explainable CED: A dataset for explainable critical error detection in machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 25–35.
- Kocmi, T. and Federmann, C. (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.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2023). Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.
- Lommel, A., Uszkoreit, H., and Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: traducció i tecnologies de la informació i la comunicació*, (12):455–463.
- O’Hagan, M. and Mangiron, C. (2013). *Game localization: Translating for the global digital entertainment industry*. John Benjamins Publishing.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.
- Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (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.
- Rei, R., Treviso, M., Guerreiro, N. M., Zerva, C., Farinha, A. C., Maroti, C., de Souza, J. G., Glushkova, T., Alves, D., Coheur, L., Lavie, A., and Martins, A. F. (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.
- Sellam, T., Das, D., and Parikh, A. P. (2020). BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.
- Valli, P. (2015). The TAUS quality dashboard. In *Proceedings of Translating and the Computer 37*, London, UK. AsLing.

A Prompt Templates

This appendix provides the complete prompt templates used in our LQA validation system. All prompts were applied consistently across all 8 evaluated models with only minor formatting adjustments for API compatibility.

A.1 System Architecture

Our LQA system uses a two-stage validation architecture:

1. **Stage 1: Rule-Based Detection** — Deterministic pattern matching flags potential issues (placeholder mismatches, encoding errors, length violations)
2. **Stage 2: LLM Validation** — LLM re-evaluates each flagged issue to determine if it is a true positive or false positive

This architecture enables high throughput by reducing LLM calls to only ambiguous cases, while maintaining high precision through LLM reasoning.

A.2 System Prompt

All 8 models use the following system prompt:

```
You are an expert localization QA analyst
with deep knowledge of translation quality
assessment. Your task is to re-evaluate
flagged issues to determine if they are
TRUE POSITIVES or FALSE POSITIVES.
Be precise, context-aware, and consider
domain-specific terminology.
```

This prompt frames the task as *validation* rather than *detection*, which improves precision by reducing false positives while maintaining high recall through the initial rule-based stage.

A.3 User Prompt Template: General Quality Check

The primary prompt used in our benchmark is the `general_quality` template (version 2.3):

```
Task: Detect translation quality issues in this
localization.
```

```
**Context:**
- Source Language: English
- Target Language: {target_lang}
- Domain: {domain}

**Text:**
- Source: "{source_text}"
- Target: "{target_text}"

**What TO Flag (CONFIRM)**:
1. Untranslated full sentences
2. Grammar errors: typos, duplications,
   malformed text
3. Missing/broken placeholders ({0}, %s, etc.)
4. Semantic errors: wrong meaning
5. Punctuation errors: excessive repetition

**What NOT TO Flag (REJECT)**:
1. Gaming/technical terms in English
   (XP, HP, MP, DPS, API, WiFi)
2. Proper nouns (brand/place/character names)
3. Cultural adaptations{lang_specific_note}

**Detection Rule: Prioritize Recall Over
Precision**
- If uncertain but suspicious -> CONFIRM
  (better false positive than missed bug)
- Only REJECT if clearly acceptable
- Aim for HIGH RECALL even if some false
  positives

Respond ONLY with valid JSON:
{{
  "decision": "CONFIRM" | "REJECT" | "UNCERTAIN",
  "confidence": 0.0-1.0,
  "reasoning": "Brief explanation",
  "category": "untranslated" | "grammar" |
    "tone" | "placeholder" |
    "particle_error" |
    "honorific_error" | "punctuation" |
    "semantic_error" | "other"
}}

{examples}

Now evaluate:
```

A.3.1 Template Variables

The prompt is dynamically populated with:

- `target_lang` — Target language code (zh, ja, ko, de, fr, es)
- `domain` — Always “gaming” for our benchmark
- `source_text` — English source string
- `target_text` — Translated string (clean or bug-seeded)
- `lang_specific_note` — Language-specific grammar notes (see below)
- `examples` — Language-specific few-shot examples (see below)

A.4 Language-Specific Adaptations

To improve cross-lingual performance, we include language-specific notes and examples.

A.4.1 Language-Specific Grammar Notes

Chinese (zh):

" : Chinese measure words (个/只/条),
punctuation (。 vs !) "

Japanese (ja):

" : Japanese particles (は/が/を),
honorifics (です/ます) "

Korean (ko):

" : Korean particles (이/가/을/를),
honorifics (요/습니다) "

German (de):

" : German cases (nominative/accusative/
dative/genitive) "

French (fr):

" : French gender agreement (le/la),
contractions (l'/d') "

Spanish (es):

" : Spanish gender agreement (el/la),
accent marks (á/é/í) "

A.4.2 Few-Shot Examples (Chinese)

Examples:

```
Source: "Click OK to continue"
Target: "Click OK to continue"
-> {"decision": "CONFIRM", "confidence": 0.9,
    "reasoning": "untranslated full sentence",
    "category": "untranslated"}

Source: "Defeat all enemies"
Target: " 打败所有敌人人"
-> {"decision": "CONFIRM", "confidence": 0.9,
    "reasoning": "character duplication ' 人人'",
    "category": "grammar"}

Source: "Cost: {{0}} gold"
Target: " 价格: 金币"
-> {"decision": "CONFIRM", "confidence": 0.95,
    "reasoning": "missing placeholder {{0}}",
    "category": "placeholder"}

Source: "Player gained {{0}} XP"
Target: " 玩家获得了 {{0}} 经验值"
-> {"decision": "REJECT", "confidence": 0.95,
    "reasoning": "correct translation,
                XP may remain",
    "category": "other"}
```

Each of the 6 languages has an equivalent 4-example set demonstrating: (1) untranslated full sentence (CONFIRM); (2) grammar error with word duplication (CONFIRM); (3) missing placeholder (CONFIRM); (4) acceptable use of English gaming term (REJECT).

A.5 Specialized Validation Templates

In addition to the general quality check, our system includes specialized templates for targeted validation:

A.5.1 Untranslated Strings Validator (v1.1)

Used when rule-based detection flags identical source and target text:

Task: Determine if this flagged "untranslated" issue is a TRUE POSITIVE or FALSE POSITIVE.

Text:

- Source: "{source_text}"

```

- Target: "{target_text}"

**Evaluation:**
CONFIRM if genuinely untranslated. REJECT if
legitimately identical (placeholders, proper
nouns, gaming/technical terms, universal terms).

**When to REJECT:**
- Placeholders only: {{0}}, %s, ${var}}
- Proper nouns: PlayStation, character/place
  names
- Universal terms: OK, WiFi, Email, USB
- Gaming terms: XP, HP, MP, DPS
- Technical terms: API, HTTP, JSON

Respond ONLY with valid JSON:
{
  "decision": "CONFIRM" | "REJECT" | "UNCERTAIN",
  "confidence": 0.0-1.0,
  "reasoning": "Brief explanation",
  "category": "placeholder" | "proper_noun" |
    "universal_term" | "gaming_term" |
    "technical_term" |
    "truly_untranslated" | "other"
}

```

A.5.2 Placeholder Error Validator (v1.1)

Used when rule-based detection flags placeholder mismatches:

```

Task: Validate if this flagged placeholder
error is a TRUE POSITIVE or FALSE POSITIVE.

**Text:**
- Source: "{source_text}"
- Target: "{target_text}"
- Issue: {issue_explanation}

**Evaluation:**
CONFIRM if placeholder missing/malformed/
wrong order. REJECT if false alarm.

Respond ONLY with valid JSON:
{
  "decision": "CONFIRM" | "REJECT" | "UNCERTAIN",
  "confidence": 0.0-1.0,
  "reasoning": "Brief explanation",
  "category": "missing_placeholder" |
    "malformed_placeholder" |
    "incorrect_order" | "false_alarm" |
    "other"
}

```

A.5.3 Grammar Error Validator (v1.1)

Used for linguistic error validation:

```

Task: Validate if this flagged grammar error
is a TRUE POSITIVE or FALSE POSITIVE.

**Text:**
- Source: "{source_text}"
- Target: "{target_text}"
- Issue: {issue_explanation}

**Evaluation:**
CONFIRM if genuine error. REJECT if intentional
(informal tone, gaming jargon, UI brevity,
dialect).

Respond ONLY with valid JSON:
{
  "decision": "CONFIRM" | "REJECT" | "UNCERTAIN",
  "confidence": 0.0-1.0,
  "reasoning": "Brief explanation",
  "category": "genuine_error" |
    "intentional_informal" |
    "domain_jargon" |
    "acceptable_variation" | "other"
}

```

A.6 Model Hyperparameters

All 8 models were evaluated with consistent hyperparameters to ensure fair comparison:

Parameter	Value
Temperature	0.1
Max Tokens	300
Top-P	0.95 (default)
Frequency Penalty	0
Presence Penalty	0

Table 8: Hyperparameters used for all 8 LLMs in benchmark evaluation.

Temperature 0.1 was chosen to minimize non-deterministic variation while preserving reasoning flexibility. Lower temperatures (0.0) produced overly rigid responses, while higher temperatures (0.3+) increased response variance and reduced reproducibility.

B Batch Size Control: Experimental Setup

This appendix provides complete experimental details for our batch size study.

B.1 Clarification: Prompt-Level Packing vs. Provider-Side Batching

Our “batch size” refers to **prompt-level packing**—concatenating multiple test samples into a single API request with structured formatting. This is *distinct from* provider-side request batching (which is not user-controllable). Prompt-level packing is a user-controllable optimization that works with any API provider.

B.2 Example: Batch Size N=3

Here is an example showing how 3 samples are packed into a single API request:

```
You are evaluating 3 game translation samples.  
For each, determine if a bug is present.
```

```
--- Sample 1 ---  
Source: "Welcome to the game"  
Translation (Chinese): " 欢迎来到游戏"  
Context: Main menu greeting  
  
--- Sample 2 ---  
Source: "{player_name} has {gold} gold"  
Translation (Chinese):  
    "{player_name} 有 {gold} 金币"  
Context: Inventory display
```

C.1 Bug Category Distribution

The examples above reflect our target distribution calibrated against industry LQA logs:

- **Linguistic errors (42%):** Mistranslation, grammar, punctuation, omission, terminology
- **Technical errors (35%):** Placeholders, encoding, formatting, UI length
- **Cultural errors (13%):** Honorifics, register, cultural references

```
--- Sample 3 ---  
Source: "Press any key to continue"  
Translation (Chinese): " 按任意键继续"  
Context: Loading screen
```

```
Provide verdicts in the format:  
Sample 1: [BUG / NO BUG]  
Sample 2: [BUG / NO BUG]  
Sample 3: [BUG / NO BUG]
```

B.3 Experimental Setup

Controlled Variables: batch sizes $N \in \{1, 2, 3, 4, 5\}$; all 8 models; 500 samples per language-genre-model setting; fixed prompt template with numbered separators; temperature 0.1.

Measured Metrics: (1) latency (end-to-end API call time; amortized per sample for $N > 1$); (2) accuracy (F1, precision, recall); (3) cost (token usage per sample).

Implementation: Python `asyncio` for concurrent API calls within each batch configuration; 3 replications per configuration to measure variance.

C Comprehensive Bug Examples

This appendix provides a comprehensive reference table of bug-seeded examples across all error categories, difficulty levels, and target languages evaluated in our benchmark.

- **Contextual errors (10%):** Wrong word choice, ambiguity, literal translation

C.2 Difficulty Level Characteristics

Easy (35%) Mechanical errors detectable by any fluent speaker: punctuation missing, word duplication, character swaps, obvious encoding errors.

Medium (46%) Language-specific errors requiring grammatical knowledge: particle errors (CJK), article gender (Romance/Germanic),

verb tense, number agreement, partial omissions.

Hard (19%) Subtle errors requiring context or cul-

tural understanding: honorific mismatches, register shifts, cultural references, contextual word choice, idiomatic vs. literal translation.

Bug Category	Source (English)	Clean Translation	Bug-Seeded Translation	Language	Difficulty
LINGUISTIC ERRORS					
Mistranslation	Defeat the dragon	Vence al dragón (es)	Come al dragón (<i>eat</i> vs <i>defeat</i>)	Spanish	Hard
Grammar (Particle)	The player earned gold	プレイヤーは金貨を獲得した (ja)	プレイヤー 金貨を獲得した (particle は missing)	Japanese	Medium
Grammar (Article)	The sword is powerful	Das Schwert ist mächtig (de)	Die Schwert ist mächtig (gender error)	German	Medium
Grammar (Agreement)	The players win	Les joueurs gagnent (fr)	Les joueurs gagne (plural subj, singular verb)	French	Medium
Verb Tense	You defeated the boss	Tu as vaincu le boss (fr, past)	Tu vaines le boss (fr, present shift)	French	Medium
Punctuation	Quest complete!	任务完成! (zh)	任务完成 (missing !)	Chinese	Easy
Word Duplication	Save the game	Sauvegarder le jeu (fr)	Sauvegarder le le jeu (doubled article)	French	Easy
Partial Omission	Collect 10 wood	收集 10 个木材 (zh)	收集 10 木材 (classifier ↑ missing)	Chinese	Medium
Terminology Swap	Restore 50HP	HP50 回復 (ja)	MP50 回復 (HP→MP swap)	Japanese	Medium
TECHNICAL ERRORS					
Placeholder Omission	{player} collected {count} items	{player} 已 收 集 个物品 (zh)	已 收集 个物品 (both placeholders removed)	Chinese	Easy
Placeholder Corruption	Welcome, {name}!	¡Bienvenido, {name}! (es)	¡Bienvenido, {nombre}! (translated place-holder)	Spanish	Medium
Encoding Error	Café menu	Menu du café (fr)	Menu du café© (mojibake UTF-8/Latin-1)	French	Easy
UI Length	Inventory full (12 char)	Inventar voll (13 ch, de)	Dein Inventar ist vollstándig belegt (38 ch)	German	Hard
Formatting	New game	新しいゲーム (ja)	新しい ゲーム (double space)	Japanese	Easy
Line Break	Start quest	Iniciar misión (es)	Iniciar\misión (line break in UI)	Spanish	Medium
CULTURAL ERRORS					
Honorific Mismatch	[Villain] Challenge me!	私に挑戦するの か! (ja, casual)	私に挑戦なさるのです か? (honorific, inappropriate)	Japanese	Hard
Register Shift	Can you help me?	¿Puedes ayudarme? (es, informal tú)	¿Puede ayudarme? (formal usted, inconsistent)	Spanish	Hard
Formality Shift	Thanks for your help	Merci pour ton aide (fr, tu)	Merci pour votre aide (fr, vous shift)	French	Hard
Cultural Reference	Travel 500 feet north	向北移动 500 英尺 (zh, feet)	向北移动 152 米 (zh, converted to meters)	Chinese	Hard
Currency Error	Costs 100 gold	100 ゴールド (ja)	100 円 (ja, yen instead of game currency)	Japanese	Medium
CONTEXTUAL ERRORS					
Wrong Word Choice	Unlock the chest	Ouvrir le coffre (fr, <i>open</i>)	Déverrouiller le coffre (fr, <i>unlock</i> literal)	French	Hard
Literal Translation	Break a leg! (good luck)	頑張って! (ja, idiom equivalent)	足を折れ! (ja, literal break leg)	Japanese	Hard

Table 9: Comprehensive bug examples across all categories (Linguistic, Technical, Cultural, Contextual) and difficulty levels (Easy, Medium, Hard) for 6 target languages. Each row shows source text, clean professional translation, and bug-seeded version with specific error type.