
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.

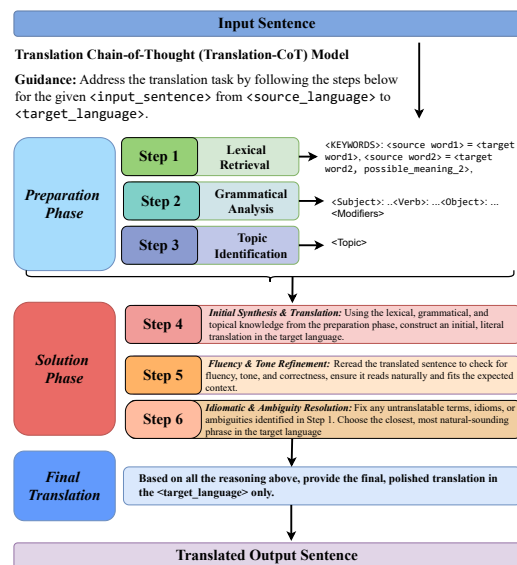


Figure 1: Architecture of the Translation-CoT framework. The framework translates source sentences into target languages in three phases: preparation, solution, and final translation. In the preparation phase, it identifies lexicons, grammatical structure, and topics in the source sentence. The solution 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.

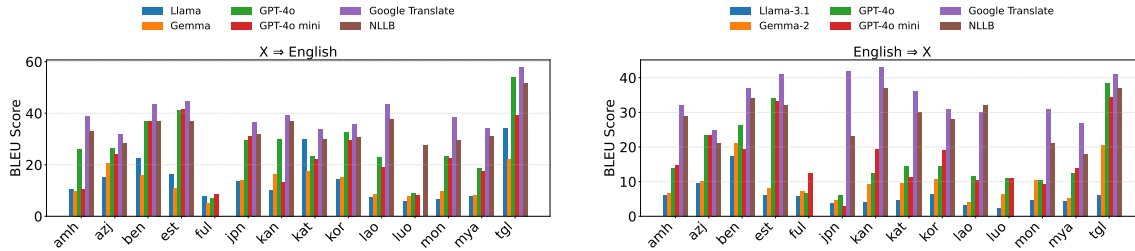


Figure 2: Translation performance (BLEU) of GPT-4o, GPT-4o-mini, Llama 3.1, and Gemma 2 across the evaluated languages. “ $X \rightarrow \text{Eng}$ ” and “ $\text{Eng} \rightarrow X$ ” denote translation into English and from English, respectively. BLEU scores are averaged over the three prompt strategies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

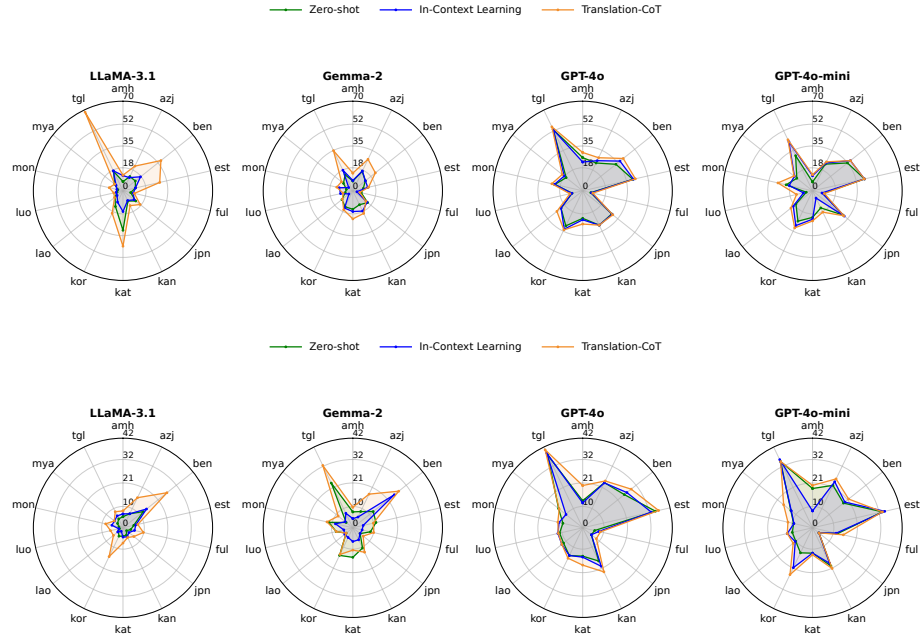


Figure 3: Multilingual translation performance (BLEU) of selected large language models (LLMs) for non-English $\rightarrow$ English (top) and English $\rightarrow$ non-English (bottom) translation across three prompt strategies: Zero-shot, In-context Learning, and Translation-CoT. In almost all cases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

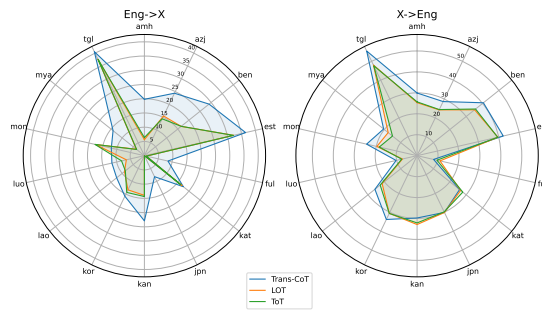


Figure 4: Comparison of CoT-based prompting strategies (Translation-CoT, ToT, and LOT) for multilingual translation performance (BLEU scores) across 14 languages from diverse language families using GPT-4o, for both translation directions (left: Eng $\rightarrow$ X, right: X $\rightarrow$ Eng).

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).

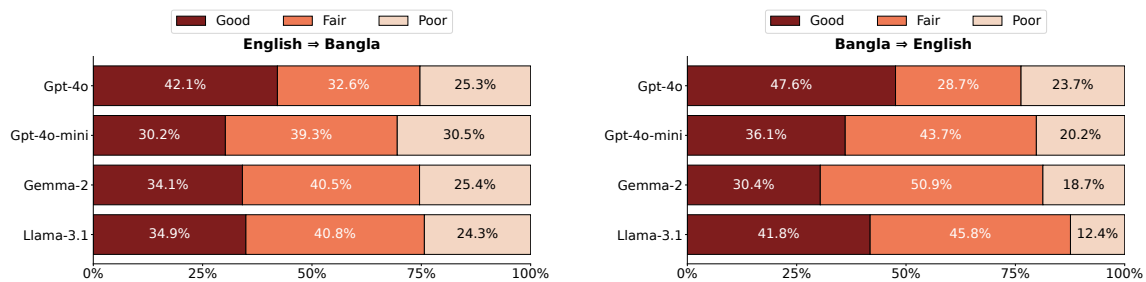


Figure 5: Human preference study comparing Translation-CoT with Zero-Shot and In-Context Learning for LLMs (GPT-4o, GPT-4o-mini, LLaMA 3.1, and Gemma 2).

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).

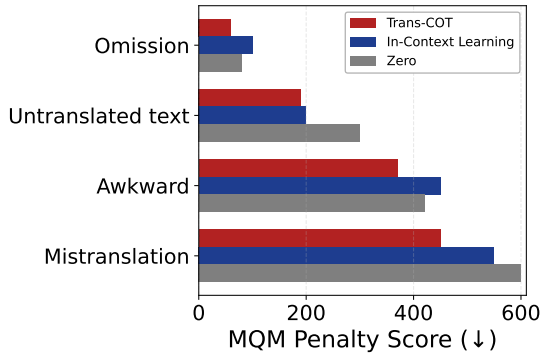


Figure 6: MQM penalty scores across different error categories for 200 test sentences from each of the English↔Bangla test sets. Lower scores indicate less severe translation errors.

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.

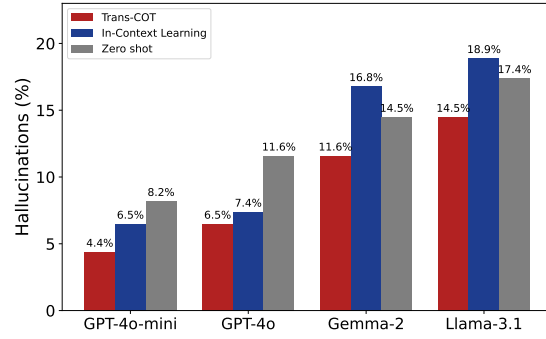


Figure 7: Ratio of hallucinations in generated translations for 200 test sentences from each of the English↔Bangla test sets. Human annotators labeled each output as either containing or not containing a hallucination error.

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., Altenschmidt, 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., Kosc, 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.

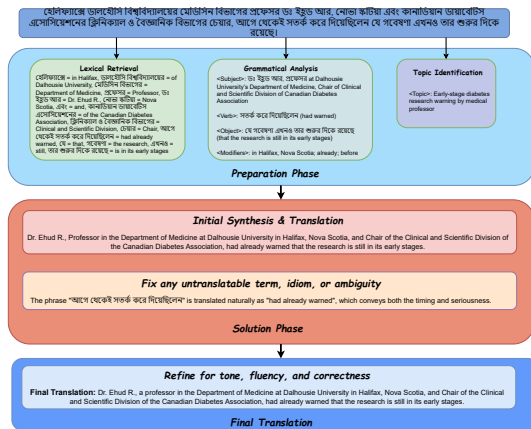


Figure 8: Example of translating a source sentence (source language: Bangla) into the target sentence (English) using the Translation-CoT reasoning steps (response from GPT-4o, accessed August 12, 2025).

C Prompt Strategies

We employ three prompting strategies for multilingual 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 $\rightarrow$ Eng and 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.

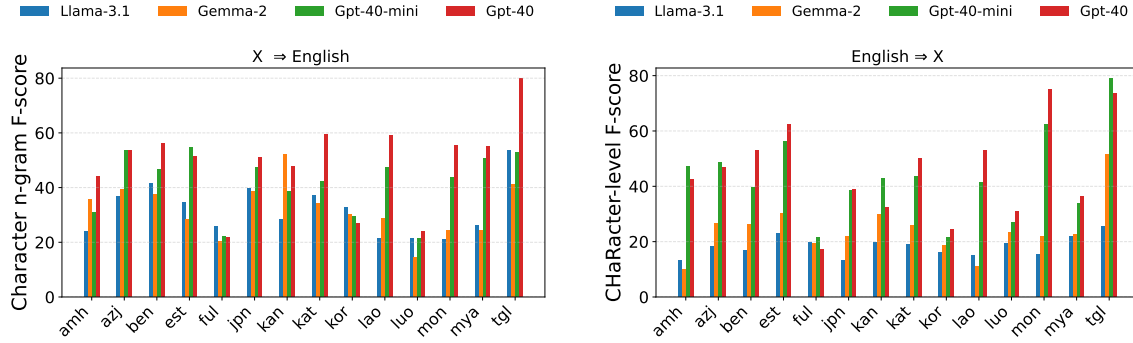


Figure 9: Translation performance (ChrF) of GPT-4o, GPT-4o-mini, LLaMA 3.1, and Gemma 2 across the evaluated languages. “X → Eng” and “Eng → X” denote translation into English and from English, respectively. ChrF scores are averaged over the three prompt strategies used in this experiment for each language.

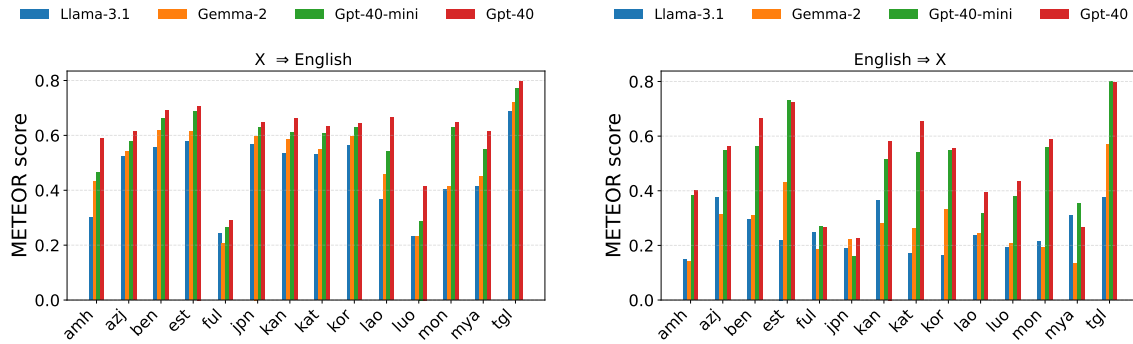


Figure 10: Translation performance (METEOR) of GPT-4o, GPT-4o-mini, LLaMA 3.1, and Gemma 2 across the evaluated languages. “X → Eng” and “Eng → X” denote translation into English and from English, respectively. METEOR scores are averaged over the three prompt strategies used in this experiment for each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

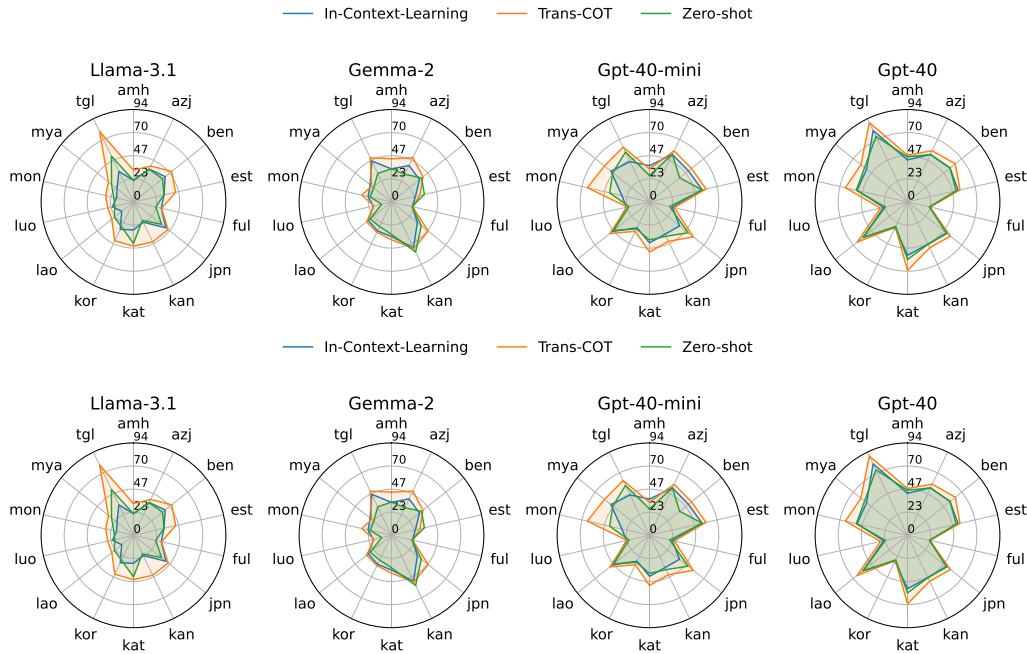


Figure 11: Multilingual translation performance (ChrF) of selected large language models (LLMs) for non-English (X) → English (top) and English → non-English (X) (bottom) translation across three prompt strategies: Zero-shot, In-context Learning, and Translation-CoT. In the majority of cases (8 out of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

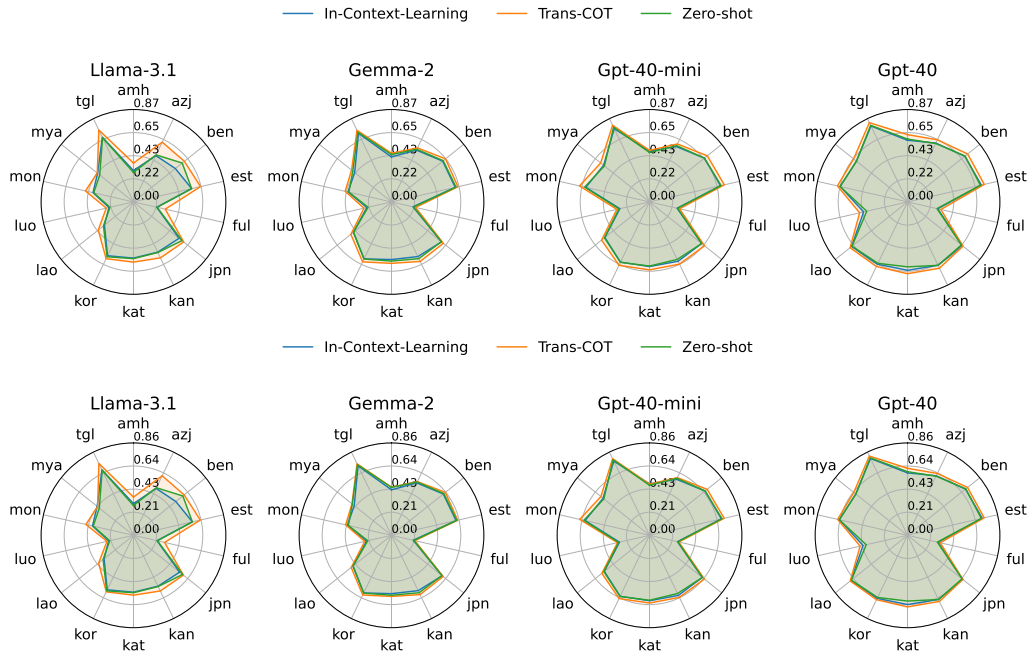


Figure 12: Multilingual translation performance (METEOR) of selected large language models (LLMs) for non-English(X) $\rightarrow$ English (top) and English $\rightarrow$ non-English (X) (bottom) translation across three prompt strategies: Zero-shot, In-context Learning, and Translation-CoT. In almost 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

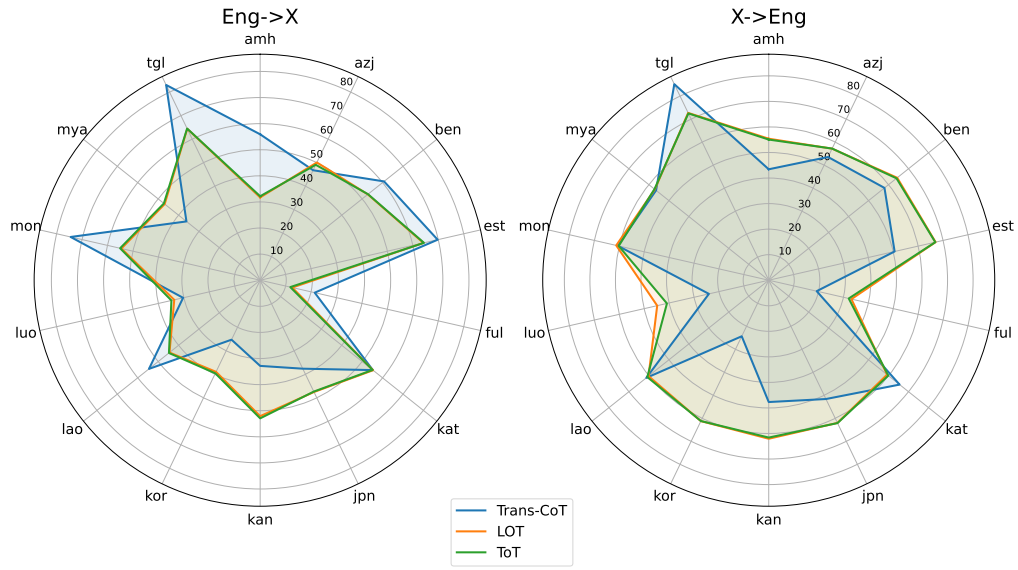
The same three bilingual annotators independently assigned each translation a binary label indicating the presence or absence of hallucination. A translation was labeled as hallucinatory if it introduced information, entities, events, attributes, or claims unsupported by or inconsistent with the source 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

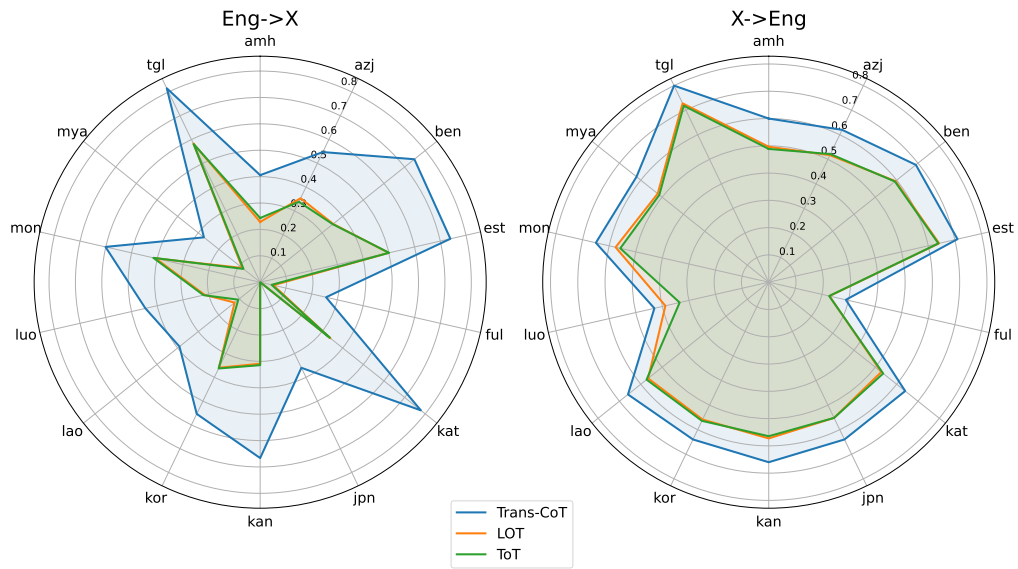
Table 3: Translation performance (BLEU, ChrF, METEOR) of GPT-4o, GPT-4o-mini, LLaMA-3.1, and Gemma-2 across 14 evaluated languages for “X→Eng” translation. Scores are the average of each test set for each language, measured using BLEU, ChrF, and METEOR metrics. Blue shading indicates that Translation-CoT 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: Eng $\rightarrow$ X, right: X $\rightarrow$ 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.