

LLM Reasoning for Machine Translation: Synthetic Data Generation over Thinking Tokens

Armel Zebaze Rachel Bawden Benoît Sagot

Inria, Paris, France

firstname.lastname@inria.fr

Abstract

Large reasoning models (LRMs) have led to new possibilities in terms of problem-solving, through the devising of a natural language thought process prior to answering a query. While their capabilities are well known across mathematics and coding tasks, their impact on the task of machine translation (MT) remains underexplored. In this work, we explore the benefits of the generation of intermediate tokens when performing MT across multiple language pairs of different levels of resourcedness and multiple setups. We first show that the “thinking tokens” of general-purpose LRMs do not help them better perform MT, motivating the question of what kind of intermediate tokens could help when building a dedicated thinking MT model via fine-tuning. We then show that distilling chains of thought (CoT) from a teacher model using CoT templates inspired by human translators’ step-by-step practices fails to outperform standard input-output fine-tuning. In contrast, constructing CoT traces with prompting strategies specifically designed for MT does yield improvements. A closer analysis reveals that these gains stem from the teacher’s translation attempts embedded in the traces, rather than from the surrounding reasoning. More broadly, our results suggest that leveraging a teacher to refine target translations or expand parallel corpora is more impactful than distilling its CoT explanations into “thinking” MT models.¹

1 Introduction

Large Language Models (LLMs) are general-purpose problem solvers (Touvron et al., 2023; OpenAI et al., 2024; Dubey et al., 2024; Kimi Team et al., 2025). Their instruction-following capabilities help them carry out a wide variety of requests provided by users via text. Research on alignment, typically through Reinforcement Learning from Human Feedback (RLHF) (Askell et al.,

2021; Bai et al., 2022; Ouyang et al., 2022; Rafailov et al., 2023; Lambert et al., 2025) has greatly contributed to improving the quality of LLMs’ outputs. Recently, a new paradigm has emerged: to train LLMs to “think” in natural language before answering a query. OpenAI o1 and o3 (OpenAI, 2024), DeepSeek-R1 (DeepSeek-AI et al., 2025), Qwen3 (Yang et al., 2025), Claude 4 (Anthropic, 2025) and Gemini 2.5 (Gemini Team et al., 2025) *inter alia* are instances of these Large Reasoning Models (LRMs). They capitalize on Reinforcement Learning (RL) to generalize the success of Chain-of-Thought (CoT) prompting (Wei et al., 2022) during training for improved safety robustness and performance. They particularly excel in reasoning-intensive tasks such as olympiad-level mathematics (AIME 2024/2025, HMMT etc.) and competition-level coding (Shi et al., 2024; Quan et al., 2025).

Labeling o1-like models as *thinking* presupposes that the intermediate tokens they produce before answering meaningfully reflect their reasoning process. Many works challenge this view. For instance, Bhambri et al. (2026) find little correlation between the correctness of final answers and the accuracy of intermediate traces, echoing earlier results by Turpin et al. (2023) showing that CoT explanations can be unfaithful. Ma et al. (2025) further demonstrate that “not thinking” can outperform “explicit reasoning” on certain challenging tasks, motivating exploration of “not thinking” in other settings. In this work, we examine the value of intermediate tokens in machine translation (MT) with LLMs and ask what forms of intermediate information are actually beneficial. MT is a particularly interesting task in this context as CoT prompting has been shown not to result in better translation than vanilla few-shot prompting (Nguyen and Xu, 2025). We show that:

- **The thinking mode of LRMs does not result in better MT outputs.** We carry out

¹<https://github.com/ArmelRandy/llm-reasoning-mt>

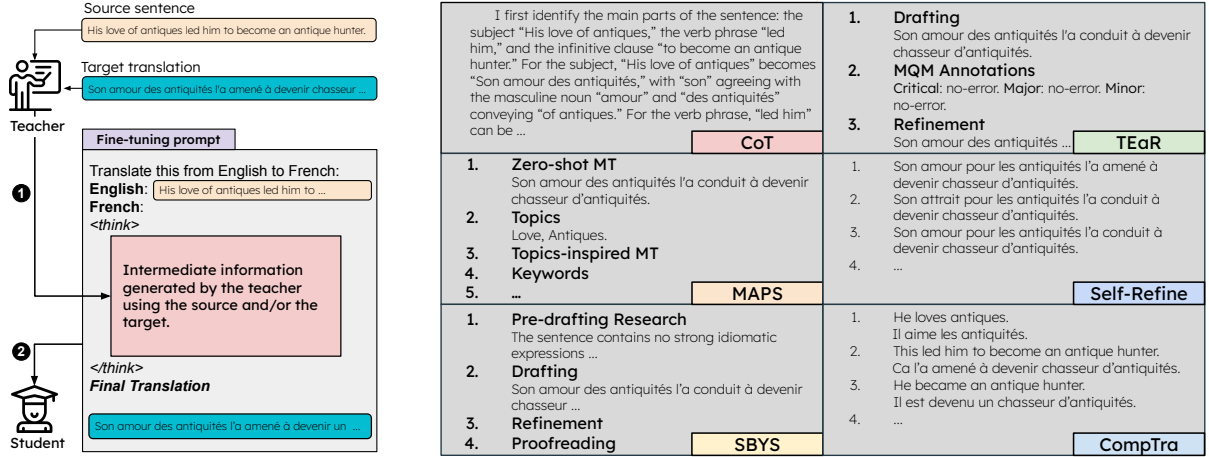


Figure 1: CoT FINE-TUNING (left): Given a source-target pair, a teacher is prompted to get a thought process on how to obtain the target given the source based on a given strategy (right). The obtained trace is used as CoT to fine-tune a student to “think” before translating.

extensive experiments across ten language directions and find no significant benefit from prior thinking. Our experiments cover zero-shot and few-shot settings, three benchmarks, high- and low-resource language pairs (X-to-English and English-to-X directions), and different temperatures of generation, and they all point to the same conclusion.

- **CoT distillation does not outperform standard fine-tuning.** Many works report a stark improvement in reasoning abilities when fine-tuning a small model to think before answering, using the CoT outputs of a teacher (Zelikman et al., 2022; Huang et al., 2024; Li et al., 2025; Guha et al., 2025). We apply this setup to fine-tune LLMs to “think” before translating and compare it to standard input-output fine-tuning. Our experiments with gemma-3-4b-pt on English to Xhosa suggest the consistent superiority of standard fine-tuning across six different MT-specific CoT templates.
- **Using traces obtained by translating the source with modular prompting strategies specifically designed for MT outperforms CoT distillation and standard input-output fine-tuning, but ultimately parallel data matters most.** Instead of vanilla CoT distillation, we propose to use the traces obtained when the teacher attempts to translate the source using a modular prompting strategy for MT. Such strategies typically involve an analysis of the source, the proposal of inter-

mediate candidates, and the derivation of the final translation. These steps can be concatenated into a single text, which we then use as intermediate information for fine-tuning the student model. This approach outperforms input-output fine-tuning by up to 3.5 BLEU (Papineni et al., 2002) points and 2 MetricX (Juraska et al., 2024) points. Analysis indicates that the gains stem mainly from translation attempts embedded in the traces. We further show that using the teacher to improve the fine-tuning dataset instead, by either enhancing the quality of its target translations or generating additional parallel pairs has greater benefits than relying on thinking tokens, without incurring extra inference cost after fine-tuning.

2 Related Work

Reasoning with LLMs. CoT prompting (Wei et al., 2022) has revolutionized the approach to reasoning with LLMs. Following In-Context Learning (ICL; Brown et al., 2020), CoT prompting drives the LLM to explain with natural language the thought process before deriving the solution to a problem. It was shown to be particularly useful for mathematical tasks that require the LLM to think through a set of reasoning steps (Cobbe et al., 2021; Hendrycks et al., 2021a,b; Suzgun et al., 2023). The intuition behind CoT prompting and its success has powered countless related prompting strategies (Kojima et al., 2022; Zhang et al., 2023; Yasunaga et al., 2024; Zebaze et al., 2024; Chen et al., 2023; Gao et al., 2023; Wang et al., 2023a;

Yao et al., 2023; Besta et al., 2024; Bi et al., 2025). CoT-based techniques have also been used to create datasets for supervised fine-tuning (Zelikman et al., 2022; Shao et al., 2024; Yue et al., 2024), which is often subsequent to prior continual pretraining on mathematics and code data (Lewkowycz et al., 2022; Azerbayev et al., 2024). We now have LRMs that are trained to produce a long chain-of-thought before responding to a query. They have set a new state of the art on multiple benchmarks (GPQA (Rein et al., 2024); SWE-Bench (Jimenez et al., 2024; Chowdhury et al., 2024)) and already motivate the creation of more challenging ones (Phan et al., 2025; Chollet et al., 2025). Despite their success, Ma et al. (2025) suggest that these thinking tokens can bring limited gains compared to not thinking on some reasoning tasks under thinking budget constraints. Others question the informativeness of CoT traces, showing that even incorrect traces can yield correct outcomes (Turpin et al., 2023; Bhambri et al., 2026) and that fine-tuning on incorrect traces can be as effective as on correct ones (Valmeekam et al., 2026). We focus on the specific task of MT and ask whether general-purpose LRMs translate better with thinking tokens or if they can be removed entirely.

Machine Translation with LLMs. Historically, encoder-decoder have been the go-to architecture for neural MT (Sutskever et al., 2014; Cho et al., 2014; Bahdanau et al., 2015; Vaswani et al., 2017). Now, decoder-only LLMs perform on par with or better than supervised MT models such as NLLB (Costa-jussà et al., 2022) on high-resource languages (HRLs), largely due to abundant high-quality web data used in pretraining. They still struggle with translating from and into low-resource languages (LRLs), but they offer more flexibility via ICL. Few-shot examples (their design, order, number and quality) greatly impact the quality of MT outputs (Agrawal et al., 2023; Moslem et al., 2023; Hendy et al., 2023; Bawden and Yvon, 2023; Mu et al., 2023; Zhu et al., 2024; Bouthors et al., 2024; Zebaze et al., 2025b). Various prompting strategies have also been developed for MT such as “Multi-Aspect Prompting and Selection” (MAPS; He et al., 2024), “Translating Step-by-Step” (SBYS; Briakou et al., 2024), “Translate, Estimate, and Refine” (TEaR; Feng et al., 2025), and “Compositional Translation” (CompTra; Zebaze et al., 2025a). This includes strategies that iteratively guide LLMs to

refine translations, with or without external feedback (Chen et al., 2024; Xu et al., 2024b; Ki and Carpuat, 2024). However, standard CoT prompting (e.g., *Let’s think step by step*) has had little to no success in LLM-based MT, with most works reporting worse results than standard input-output prompting (Peng et al., 2023; Zebaze et al., 2025a; Nguyen and Xu, 2025). Several works have explored building LRMs for MT (Wang et al., 2025a, 2026, 2025b; He et al., 2025) by closely following what is done for reasoning tasks. They typically prompt a large model (e.g., DeepSeek-R1) with a curated CoT prompt that guides it from the source sentence to the target translation, then use the generated CoT for supervised fine-tuning (SFT) followed by RL fine-tuning. However, Zheng et al. (2025) suggest that thinking does not help MT performance when applying GRPO (Shao et al., 2024) on rewards that only evaluate the final translation. In this work, we focus on SFT and study the pertinence of thinking tokens via CoT distillation, which is already successful on reasoning tasks (Huang et al., 2024; Li et al., 2025; Guha et al., 2025). Moreover, given a source, we propose to use modular prompting strategies for MT that have been shown to outperform zero- and/or few-shot MT and ask a teacher to translate the source with the strategies. The outputs (or traces) of their individual steps can be concatenated into a single text and used as intermediate tokens to fine-tune a student, thereby building a “thinking” MT model that “thinks” before translating. Rajaei et al. (2026) build on an earlier version of this work to investigate whether reasoning improves MT with LRMs, and confirm our findings. However, unlike our study, they do not disentangle the effect of translation attempts appearing within the reasoning traces.

3 Benchmarking LRMs at Scale: To Think, or Not To Think?

We first investigate the influence of reasoning on the translation quality of general-purpose LRMs. We compare two conditions: (i) the model is allowed to generate reasoning tokens prior to producing the translation (up to 3,500 tokens), and (ii) reasoning is explicitly suppressed by appending `<think>\n\n</think>` to the prompt (Non-Thinking Mode; Yang et al., 2025). We carry out experiments with the Qwen3 model family (Yang et al., 2025), ranging in size from 0.6B to 32B parameters, in a zero-shot English-to-X set-

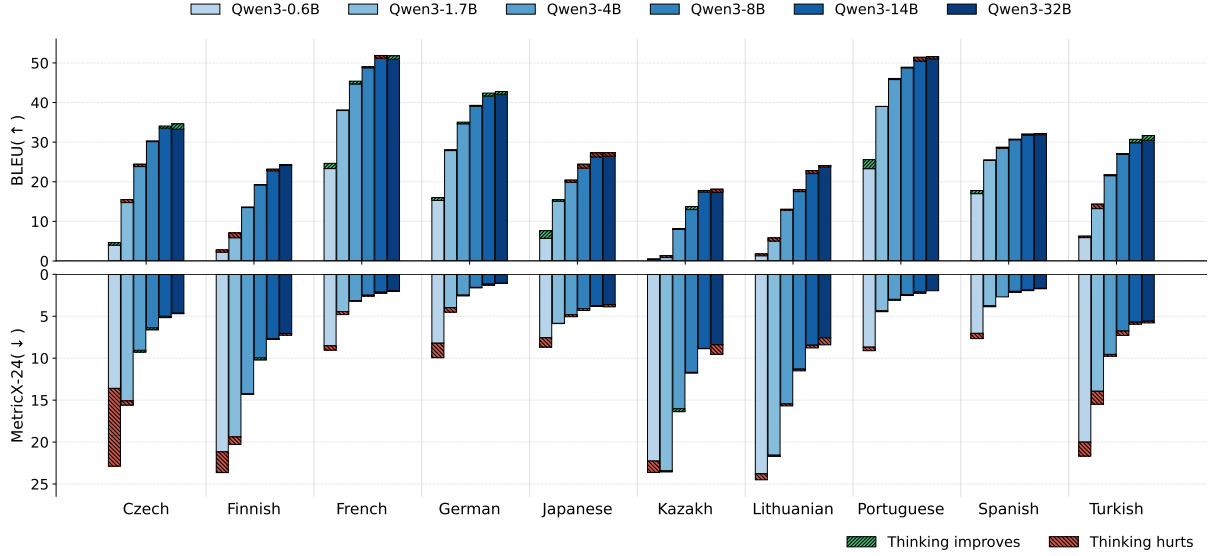


Figure 2: BLEU and MetricX scores for ten English $\rightarrow$ X directions from FLORES 200, with and without thinking. See Table 2 for numerical values.

ting² for ten FLORES-200 (Goyal et al., 2022; Costa-jussà et al., 2022) languages: Czech, Finnish, French, German, Japanese, Kazakh, Lithuanian, Portuguese, Spanish, and Turkish. The results (Figure 2) show that the performance with and without prior thinking is similar. Non-thinking is slightly better, in particular in terms of MetricX but the difference is usually less than 0.5 MetricX points. We provide additional results with more models and directions in Appendix B, and with two other benchmarks, NTREX 128 (Federmann et al., 2022; Barrault et al., 2019) and TICO-19 (Anastasopoulos et al., 2020), in Appendix B.3. The findings are consistent throughout: regardless of whether we translate into or out of English, target high-resource or low-resource languages, use alternative benchmarks, or switch between Qwen3 and DeepSeek-R1, the reasoning tokens of LRMs bring little to no benefit in terms of translation quality.

We evaluate Qwen3-32B when sampling with temperatures 0.2, 0.4, 0.6, 0.8 and 1.0 and plot the results in Figure 3. As opposed to default usage recommendations for LRMs, sampling degrades the performance, confirming the results obtained by Chen et al. (2025) with DeepSeek-R1-670B. In addition, at each temperature, outputting thinking tokens or not gives approximately the same level of performance. We observe a slight gain in BLEU (≤ 1 point), but a degradation in MetricX. Low temperatures correlate with higher performance—

in particular, $T = 0$ works best in general. These results reinforce the fact that MT does not significantly benefit from the presence of thinking tokens, which is further supported by our human evaluation (Appendix B.2).

4 Approaches to Improving MT with Intermediate Reasoning

Given that general-purpose LRMs do not seem to benefit from outputting thinking tokens prior to translation, we investigate how to build a successful “thinking” MT model, i.e., one that first produces intermediate reasoning before translation and outperforms models trained without intermediate steps. We apply COT FINE-TUNING (COTFT), whereby a student model is trained to first produce intermediate tokens (CoT or “thoughts”) before generating the target translation, as shown in Figure 1. Given a parallel dataset, we explore multiple types of intermediate information a teacher model can generate to help train a “thinking” (student) MT model (see the right side of Figure 1), which we categorize into two types: (i) CoT prompting and (ii) modular translation-specific prompting strategies.

- **CoT prompting** corresponds to the standard CoT distillation approach inherited from reasoning tasks (Figure 1, right, first box). We use the six “CoT instance construction” prompt templates reflecting human translators’ reasoning proposed by He et al. (2025) for generating cold-start data for their R1-T1 model: hierarchical trans-

²We use greedy decoding.

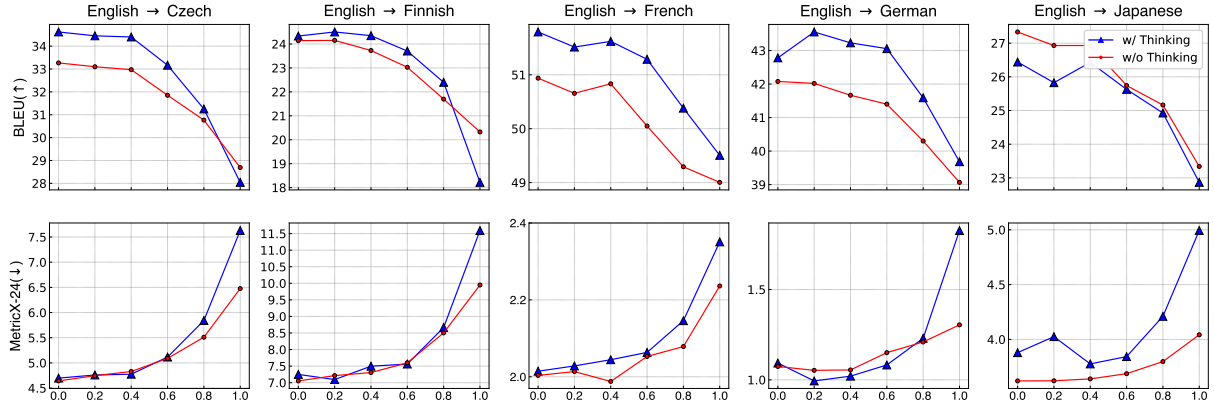


Figure 3: Impact of the temperature on the translation quality with and without thinking tokens.

lation, triangulating translation, backtranslation, context-aware translation, translation explanation and structural transformation (see Appendix A.3 for the prompt templates). For each source–target pair and a CoT template, the teacher produces a first-person thought process in which it explains how it analyzes the sentence—identifying elements such as the subject, verb, and object—and how it arrives at the target translation by reasoning about linguistic aspects (syntactic rules, word order) and the broader context.

- Alternatively, we propose using as intermediate information the traces obtained after applying **modular translation-specific prompting strategies** to translate the source. They generally involve multiple steps, (see the five other boxes of Figure 1):
 - **MAPS**: a modular process comprising source analysis (extraction of keywords, topics, and relevant demonstrations) and translation attempts (inspired by the extracted information), complemented by zero-shot translation.
 - **SBYS**: a four-step process comprising pre-drafting research (identification of expressions that may pose a challenge for translation), drafting, refinement, and proofreading (for terminology, fluency, etc.).
 - **TEaR**: a three-step process comprising translation, annotation (of potential translation errors), and annotation-based refinement.
 - **Self-Refine**: an iterative process comprising the initial translation (in a zero-shot setting) and successive rounds of self-refinement (to improve accuracy and fluency).

- **CompTra**: a three-step process comprising decomposition of the source into simpler phrases, few-shot translation of each phrase, and recombination³ (into a final output).

Given a source sentence, we apply all steps of each selected prompting strategy and concatenate the outputs to form a text, which serves as intermediate information for COTFT. We aim to see whether this approach results in improved MT models, analogous to how CoT prompting improve over direct IO prompting with standard LLMs.

We compare COTFT to INPUT-OUTPUT FINE-TUNING (IOFT), the baseline approach where the student is trained to directly predict the translation given a source sentence. In both cases, the source and target are the same; the difference lies only in the presence of intermediate reasoning during training. We provide example training samples for each strategy tested in Appendix C.2. In summary, we first examine whether prompting a teacher to emulate a human translator and produce CoT traces helps to train a thinking MT model (student) that is more effective than a standard IOFT model. Finally, we evaluate whether MT-specific prompting strategies produce higher-quality traces for training thinking MT models and discuss the broader implications.

5 Experiments

5.1 Experimental Setup

Evaluation datasets We use FLORES-200 (Goyal et al., 2022; Costa-jussà et al., 2022) devtest (1012 examples). For fine-tuning

³We do not use the output of the recombination step when building the intermediate tokens.

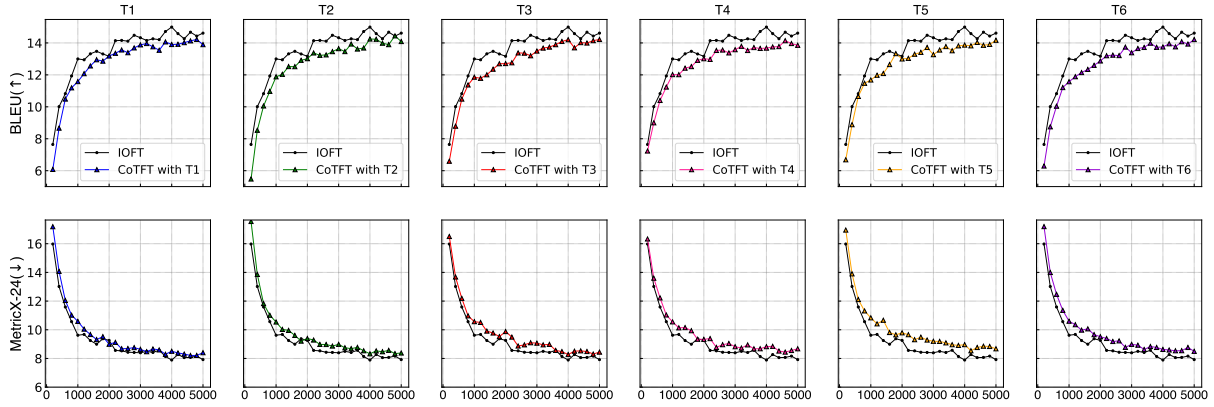


Figure 4: Comparison between IOFT and CoTFT with six different CoT templates. Across all figures, the x-axis represents the number of training steps.

experiments, we focus on the LRL Xhosa and include results for the HRL Lithuanian in Appendix B.4-B.6.

Fine-tuning datasets For Xhosa, we use Llama-4-Scout-17B (AI at Meta, 2025) and synthetic, multi-domain sentence-level data generated using the TOPXGEN pipeline (Zebaze et al., 2025c). For Lithuanian, we use the WMT19 dataset (Barrault et al., 2019) and run the same experiments as Xhosa, as detailed in Appendices B.4 to B.6.

Models For Xhosa, we use Llama-4-Scout-17B (AI at Meta, 2025) as the teacher (to generate reasoning traces for CoTFT) and gemma-3-4b-pt as the student. For Lithuanian, we pair gemma-3-27b-it as teacher with gemma-3-1b-pt as the student in Appendices B.4 to B.6.

Evaluation metrics We mainly use MetricX-24-Hybrid-XXL (Juraska et al., 2024), a reference-based metric that supports the same 101 languages as mT5 (Xue et al., 2021). It assigns a score from 0 to 25, with higher scores indicating more errors. We also use BLEU⁴ (Papineni et al., 2002) as implemented in sacreBLEU (Post, 2018).

Implementation details We fine-tune all models for 5k steps on one H100 80G with a learning rate of 1e-5, a constant scheduler with 500 warm-up steps (from 1e-6) and a batch size of 4. We use 4 gradient accumulation steps and a maximum sequence length of 512 for IOFT (16 and 2048 respectively for CoTFT). Unless stated otherwise, models are evaluated in a zero-shot fashion with

greedy decoding. See Appendices A.1 and A.2 for additional details and Appendix C.1 for ablations.

5.2 Distilled CoT as intermediate tokens

We compare IOFT and CoTFT in the CoT distillation setup. We fine-tune gemma-3-4b-pt and compare the performance of all six CoTFT variants against IOFT. The results (reported every 200 steps) in Figure 4 show that CoTFT consistently fails to improve over IOFT (in black) across all templates. The variability of performance across templates is negligible; they fall short compared to IOFT by about 0.5 BLEU and 0.5 MetricX points. We reach the same conclusions with DeepSeek-R1-Llama as the teacher (Appendix B.7).

5.3 MT traces generated by prompting strategies as intermediate tokens

Given a teacher and a modular prompting strategy, can the traces generated during translation, when used as intermediate information, help produce better outputs? As shown in Figure 5, CoTFT on traces based on such strategies outperforms IOFT, e.g. we get gains up to +3.5 BLEU and -2.0 MetricX with MAPS. For the other prompting strategies, improvements remain around +2 BLEU and -1.5 MetricX. Using CoT traces derived from these strategies appears beneficial—but why? The key difference is that, unlike pure CoT prompting, most of these strategies (except CompTra) include one or multiple drafting phases. The success of CoTFT may therefore stem from drafts that surpass the ground truth. We test this hypothesis as follows. For each strategy, we use the quality estimation score BLASER 2.0-QE (Duquenne et al.,

⁴nrefs:1|case:mixed|eff:no|tok:flores200|smooth:exp|version:2.4.2

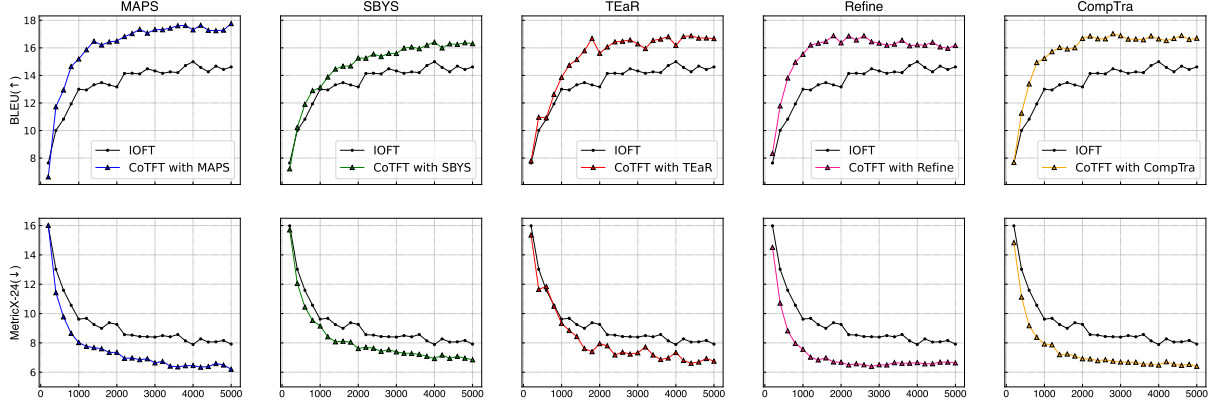


Figure 5: Comparison between IOFT and CoTFT with five different prompting strategies.

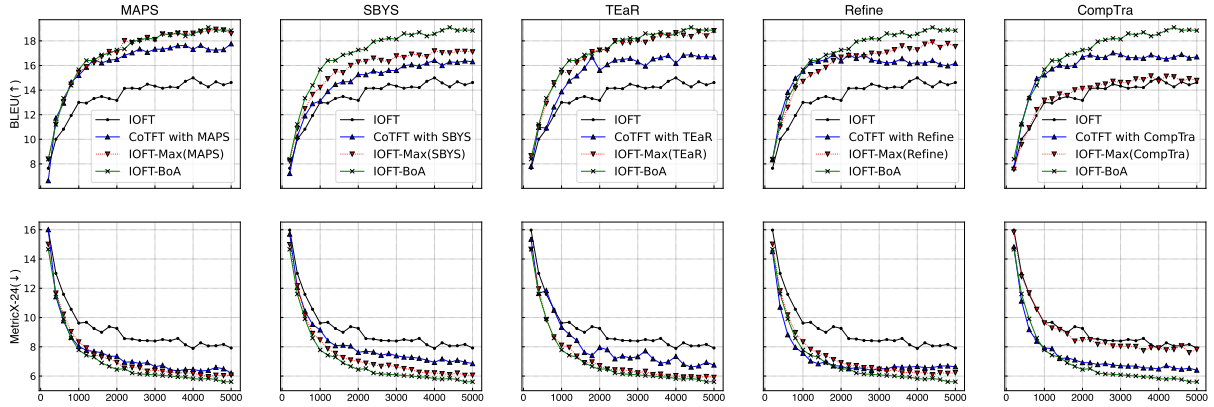


Figure 6: Comparison between IOFT, CoTFT and IOFT-MAX with five different prompting strategies.

2023; Dale and Costa-jussà, 2024) to obtain the best translation between the ground truth and the attempts embedded in the teacher’s traces. We consider 2 scenarios: (i) IOFT-MAX(STRATEGY), which is IOFT where the target is replaced by the best one between the ground truth and those potentially generated by the prompting strategy, and (ii) CoTFT-MAX(STRATEGY) which is analogous to IOFT-MAX(STRATEGY) but with the intermediate tokens. In addition to the above scenarios, we consider IOFT-BoA (best of all) which is IOFT where the target is the best between the ground truth and translations embedded into the traces obtained across all prompting strategies considered.

IOFT-MAX For MAPS, SBYS, TEaR and Self-Refine, IOFT-MAX (in red) works better than CoTFT (i.e. with the traces) and IOFT (Figure 6). This indicates that the quality of the target is an important factor for MT performance. Using better ground truths (IOFT-BoA, in green) can make IOFT go from 14 to 18 BLEU (8 to 5.6 MetricX) with the same number of parallel pairs and the same

training recipe. CompTra behaves differently; the traces of CompTra contain translations of shorter parts of the sentence built by splitting the source (not of the source itself) and their translation is unlikely to be better than the ground truth. This explains the performance similarity between standard IOFT and IOFT-MAX(COMPTRA). CoTFT with CompTra outperforms IOFT-MAX(COMPTRA) and IOFT indicating that CoTFT can be successful without including better translation attempts than the ground truth; partial translations are enough.

CoTFT-MAX For MAPS, SBYS, TEaR and Self-Refine, IOFT-MAX generally works better than CoTFT-MAX (Figure 7). This confirms the previous conclusion on these strategies, i.e. when the traces provided by the teacher do not contain translation attempts better than the ground truth, they do not help improve the MT performance. CoTFT-MAX(COMPTRA) works slightly better than IOFT-MAX(COMPTRA), but both underperform CoTFT with CompTra. This reinforces the idea that subsentences or propositions

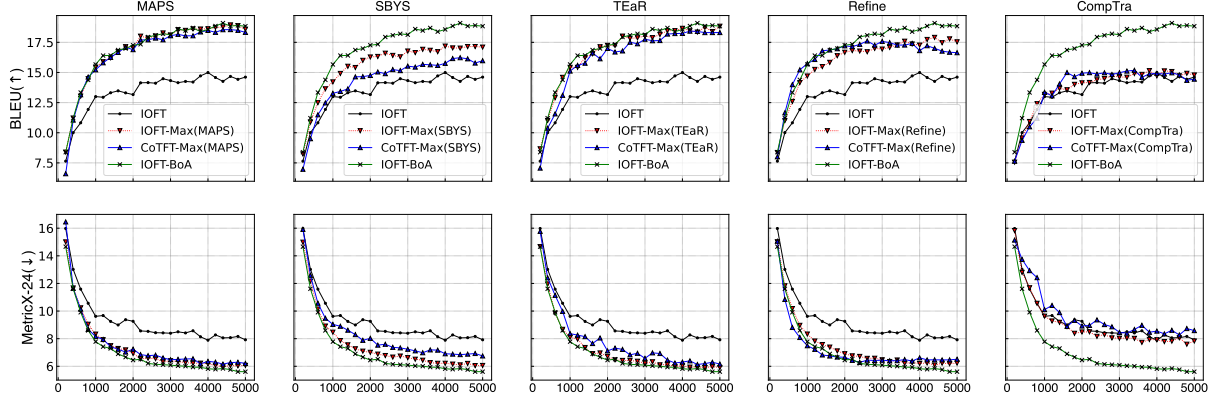


Figure 7: Comparison between IOFT, IOFT-MAX and CoTFT-MAX with five different prompting strategies.

and their translations can serve as valuable intermediate information for CoTFT. IOFT-BoA being consistently above all the curves suggests that the quality of the target translations matters more than traces. IOFT-BoA outperforms all the CoTFT approaches while being cheaper and faster to perform. We reach the same conclusions with `gemma-3-27b-it` as the teacher (Appendix B.8).

5.4 Derived sentences as intermediate tokens

We further investigate the use of alternative sentences derived from the source, along with their translations, as intermediate tokens. In CompTra, these are obtained by decomposing the source into independent propositions (Zebaze et al., 2025a). We consider three other derivation strategies: *Paraphrases* (P), *Syntactic Paraphrases* (SP) and *Hard Expressions* (H). P prompts the teacher to generate five paraphrases of the source. SP generates five sentences with the same syntax as the source (grammatical roles, syntactic dependencies etc.) but using different words. Finally, H asks the teacher to extract expressions it deems difficult to translate. In all cases, the teacher translates the expressions generated after decomposition. For each decomposition strategy (P , SP , H , and CompTra), we compare CoTFT (with the teacher’s sentence–translation pairs as reasoning traces) with IOFT. We also evaluate IOFT-EXT(*strategy*), which applies IOFT on the original dataset augmented with the generated pairs as additional training samples. As shown in Figure 8, CoTFT consistently outperforms IOFT across all decomposition strategies. CompTra and SP are the best approaches for CoTFT. IOFT-EXT(P) and IOFT-EXT(SP) result in significant gains over the IOFT baseline (+4 BLEU, -2 MetricX). IOFT-EXT also outperforms CoTFT with

P and SP . However, it has less success with H and CompTra. We attribute this to the fact that the pairs generated by H and CompTra are shorter and largely overlap with the original training samples, giving fewer gains as additional data compared to entirely new sentences. However, these short phrases (CompTra and H) are valuable intermediate information, as CoTFT outperforms IOFT and IOFT-EXT in these scenarios. IOFT-EXT(SP) and IOFT-EXT(P) are the best overall, showing the large impact of the amount of parallel data.

6 Conclusion

We explored fine-tuning LLMs to generate intermediate tokens as a method to improve their MT capabilities. Through a broad spectrum of experiments, we find that outputting reasoning traces does not help models to produce better translations (for LRMs and during CoT distillation). We also investigated how traces produced by alternative modular MT prompting strategies could help and found that traces incorporating translation attempts can serve as valuable intermediate information (Section 5.3). However, ultimately two factors affect the success of MT fine-tuning: the quality of the target translations and the quantity of parallel data. When they are both ensured, standard IOFT is enough. These findings connect to two important results in MT. First, CoT prompting has been shown to fail to improve over standard IO prompting for MT in zero-shot settings with standard LLMs (Peng et al., 2023; Nguyen and Xu, 2025). With standard LLMs, CoT prompting considerably degrades MT performance compared to IO prompting. Thinking models do not exhibit this degradation, but they do not yield improvements either (Section 3). We hypothesize that this difference stems from the fact

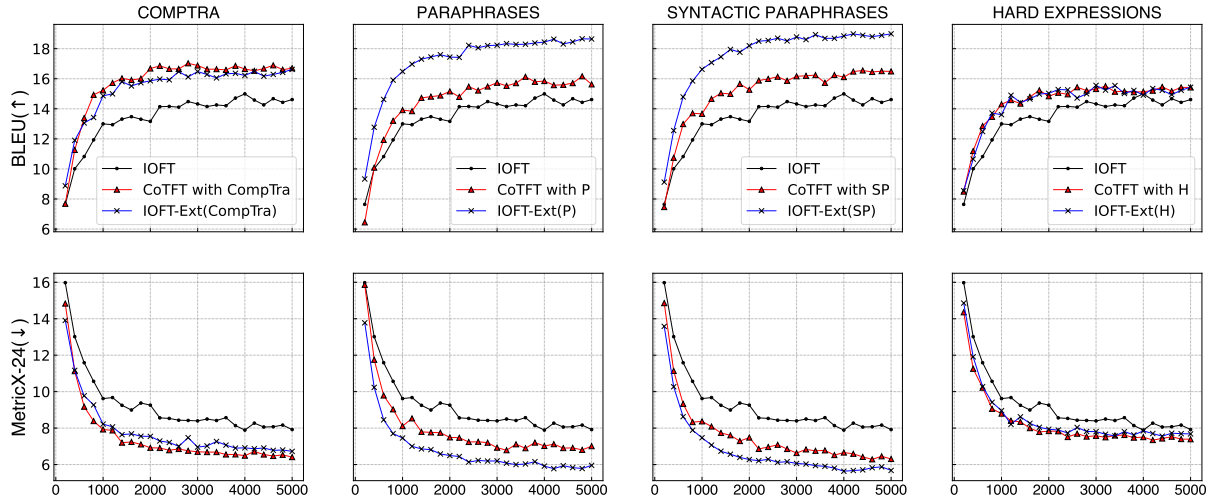


Figure 8: Comparison between IOFT and CoTFT with four sentence derivation strategies.

that thinking models are explicitly trained to produce reasoning tokens, whereas standard LLMs generate intermediate tokens that are unhelpful or even deleterious for MT. Section 5.2 further shows that generating human-inspired reasoning traces for MT using LLMs produces intermediate steps that do not enable CoTFT to surpass IOFT. We stress that this does not imply that no human-inspired strategy could work; finding one could require an entire study. MT behaves differently from typical reasoning tasks: it is far less obvious what a useful reasoning trace for translation should look like than, say, for mathematical problem solving. Our empirical results with the intuitive ways we tried to operationalize did not provide much gain. Second, our findings echo recent work on grammar books and dictionaries as external resources for MT. [Aycock et al. \(2025\)](#) showed that the parallel sentences embedded within grammar books may play a central role in the success of these approaches, though they may not be the sole contributing factor. While our work does not involve grammatical explanations, a parallel can be drawn: intermediate tokens provided no benefit when translation attempts (full or partial) were absent, but accounted for all improvements when they were present. It is worth noting that [Marmonier et al. \(2025\)](#) later showed through a set of controlled experiments that grammatical explanations can genuinely help MT, suggesting that the picture is more nuanced than bitexts alone being responsible for the gains. However, in the scenarios we tested, data quantity and quality appear to be the most decisive factors: a polished translation dataset without additional

reasoning tokens goes a long way.

Limitations

This study focuses on sentence-level MT and naturally opens an avenue for future works on more complex translation scenarios such as document-level MT and stylized MT. According to [Wang et al. \(2023b\)](#), stylized machine translation refers to translating the source sentence into different styles with certain attributes. We provide preliminary results in Appendix B.11 that seem to extend our findings to these scenarios, but we believe a more fine-grained study could provide further insights.

Acknowledgments

This work was partly funded by Benoît Sagot’s chair in PRAIRIE-PSAI, funded by the French national agency ANR, as part of the “France 2030” strategy under the reference ANR-23-IACL-0008. It was also partly funded by the French *Agence Nationale de la Recherche* (ANR) under the project TraLaLaM (“ANR-23-IAS1-0006”). This work was granted access to the HPC resources of IDRIS under the allocation 2025-AD011015933 made by GENCI.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. [In-context examples selection for machine translation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

- AI at Meta. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>.
- Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. **TICO-19: the Translation Initiative for COvid-19**. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.
- Anthropic. 2025. Introducing Claude 4. <https://www.anthropic.com/news/claude-4>.
- Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, and 3 others. 2021. **A General Language Assistant as a Laboratory for Alignment**. *Preprint*, arXiv:2112.00861.
- Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2025. **Can LLMs really learn to translate a low-resource language from one grammar book?** In *The Thirteenth International Conference on Learning Representations*.
- Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen Marcus McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2024. **Llemma: An Open Language Model for Mathematics**. In *The Twelfth International Conference on Learning Representations*.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. **Neural Machine Translation by Jointly Learning to Align and Translate**. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. **Training a helpful and harmless assistant with reinforcement learning from human feedback**. *Preprint*, arXiv:2204.05862.
- Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. **Findings of the 2019 Conference on Machine Translation (WMT19)**. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.
- Rachel Bawden and François Yvon. 2023. **Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM**. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michał Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoeffler. 2024. **Graph of Thoughts: Solving Elaborate Problems with Large Language Models**. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.
- Siddhant Bhambri, Upasana Biswas, and Subbarao Kambhampati. 2026. **Interpretable traces, unexpected outcomes: Investigating the disconnect in trace-based knowledge distillation**. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36390–36408, San Diego, California, United States. Association for Computational Linguistics.
- Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. 2025. **Forest-of-Thought: Scaling Test-Time Compute for Enhancing LLM Reasoning**. In *Forty-second International Conference on Machine Learning*, Vancouver, Canada.
- Maxime Bouthors, Josep Crego, and François Yvon. 2024. **Retrieving examples from memory for retrieval augmented neural machine translation: A systematic comparison**. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3022–3039, Mexico City, Mexico. Association for Computational Linguistics.
- Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. **Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts**. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. **Language models are few-shot learners**. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. **WIT3: Web inventory of transcribed and translated talks**. In *Proceedings of the 16th Annual Conference of the European Association for Machine*

- Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.
- Andong Chen, Yuchen Song, Wenxin Zhu, Kehai Chen, Muyun Yang, Tiejun Zhao, and Min zhang. 2025. [Evaluating o1-Like LLMs: Unlocking Reasoning for Translation through Comprehensive Analysis](#). *Preprint*, arXiv:2502.11544.
- Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. [Iterative translation refinement with large language models](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190, Sheffield, UK. European Association for Machine Translation (EAMT).
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. 2023. [Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks](#). *Transactions on Machine Learning Research*.
- Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. [Learning phrase representations using RNN encoder–decoder for statistical machine translation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.
- François Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. 2025. [ARC-AGI-2: A New Challenge for Frontier AI Reasoning Systems](#). *Preprint*, arXiv:2505.11831.
- Neil Chowdhury, James Aung, Chan Jun Shern, Oliver Jaffe, Dane Sherburn, Giulio Starace, Evan Mays, Rachel Dias, Marwan Aljubei, Mia Glaese, Carlos E Jimenez, John Yang, Kevin Liu, and Aleksander Madry. 2024. Introducing SWE-bench Verified. *OpenAI*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Mailhard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. [No Language Left Behind: Scaling Human-Centered Machine Translation](#). *CoRR*, abs/2207.04672.
- David Dale and Marta R. Costa-jussà. 2024. [BLASER 2.0: a metric for evaluation and quality estimation of massively multilingual speech and text translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16075–16085, Miami, Florida, USA. Association for Computational Linguistics.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. [DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning](#). *Preprint*, arXiv:2501.12948.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 514 others. 2024. [The Llama 3 Herd of Models](#). *Preprint*, arXiv:2407.21783.
- Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. [SONAR: Sentence-Level Multimodal and Language-Agnostic Representations](#). *Preprint*, arXiv:2308.11466.
- Christian Federmann, Tom Kocmi, and Ying Xin. 2022. [NTREX-128 – News Test References for MT Evaluation of 128 Languages](#). In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.
- Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2025. [TEaR: Improving LLM-based machine translation with systematic self-refinement](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3922–3938, Albuquerque, New Mexico. Association for Computational Linguistics.
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. [PAL: Program-aided language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.
- Gemini Team et al. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. [The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation](#). *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins.

2024. [xcomet: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. <https://github.com/huggingface/accelerate>.
- Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, and 31 others. 2025. [OpenThoughts: Data Recipes for Reasoning Models](#). *Preprint*, arXiv:2506.04178.
- Mingui He, Yilun Liu, Shimin Tao, Yuanchang Luo, Hongyong Zeng, Chang Su, Li Zhang, Hongxia Ma, Daimeng Wei, Weibin Meng, Hao Yang, Boxing Chen, and Osamu Yoshie. 2025. [R1-T1: Fully Incentivizing Translation Capability in LLMs via Reasoning Learning](#). *Preprint*, arXiv:2502.19735.
- Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. [Exploring human-like translation strategy with large language models](#). *Transactions of the Association for Computational Linguistics*, 11:229–246.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. [Measuring Massive Multitask Language Understanding](#). In *International Conference on Learning Representations*, Vienna, Austria.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. [Measuring Mathematical Problem Solving With the MATH Dataset](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. [How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation](#). *Preprint*, arXiv:2302.09210.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-Rank Adaptation of Large Language Models](#). In *International Conference on Learning Representations*, Virtual.
- Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. 2024. [O1 Replication Journey – Part 2: Surpassing O1-preview through Simple Distillation, Big Progress or Bitter Lesson?](#) *Preprint*, arXiv:2411.16489.
- Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. [SWE-bench: Can language models resolve real-world github issues?](#) In *The Twelfth International Conference on Learning Representations*, Vienna, Austria.
- Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. [MetricX-24: The Google submission to the WMT 2024 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.
- Dayeon Ki and Marine Carpuat. 2024. [Guiding Large Language Models to Post-Edit Machine Translation with Error Annotations](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4253–4273, Mexico City, Mexico. Association for Computational Linguistics.
- Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijie Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chen-zhuang Du, Dikang Du, Yulun Du, Yu Fan, and 150 others. 2025. [Kimi K2: Open Agentic Intelligence](#). *Preprint*, arXiv:2507.20534.
- Philipp Koehn. 2004. [Statistical significance tests for machine translation evaluation](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient Memory Management for Large Language Model Serving with PagedAttention](#). In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. [Tulu 3: Pushing Frontiers in Open Language Model Post-Training](#). *Preprint*, arXiv:2411.15124.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. [Solving Quantitative Reasoning Problems with Language Models](#). In

- Advances in Neural Information Processing Systems*, volume 35, pages 3843–3857. Curran Associates, Inc.
- Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. [Language models can easily learn to reason from demonstrations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15979–15997, Suzhou, China. Association for Computational Linguistics.
- Xing Han Lù. 2024. [BM25S: Orders of magnitude faster lexical search via eager sparse scoring](#). *Preprint*, arXiv:2407.03618.
- Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025. [Reasoning Models Can Be Effective Without Thinking](#). *Preprint*, arXiv:2504.09858.
- Malik Marmonier, Rachel Bawden, and Benoît Sagot. 2025. [Explicit learning and the LLM in machine translation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31372–31422, Suzhou, China. Association for Computational Linguistics.
- Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. [Adaptive machine translation with large language models](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.
- Yongyu Mu, Abudurexiti Reheman, Zhiqian Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. [Augmenting large language model translators via translation memories](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10287–10299, Toronto, Canada. Association for Computational Linguistics.
- Lam Nguyen and Yang Xu. 2025. [Reasoning for translation: Comparative analysis of chain-of-thought and tree-of-thought prompting for LLM translation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 259–275, Vienna, Austria. Association for Computational Linguistics.
- OpenAI. 2024. [Learning to reason with LLMs](#).
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [GPT-4 Technical Report](#). *Preprint*, arXiv:2303.08774.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, page 311–318, USA. Association for Computational Linguistics.
- Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. [Towards making the most of ChatGPT for machine translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.
- Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, and 1090 others. 2025. [Humanity’s last exam](#). *Preprint*, arXiv:2501.14249.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Shanghaoran Quan, Jiaxi Yang, Bowen Yu, Bo Zheng, Dayiheng Liu, An Yang, Xuancheng Ren, Bofei Gao, Yibo Miao, Yunlong Feng, Zekun Wang, Jian Yang, Zeyu Cui, Yang Fan, Yichang Zhang, Binyuan Hui, and Junyang Lin. 2025. [CodeElo: Benchmarking Competition-level Code Generation of LLMs with Human-comparable Elo Ratings](#). *Preprint*, arXiv:2501.01257.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. [Direct Preference Optimization: Your Language Model is Secretly a Reward Model](#). In *Advances in Neural Information Processing Systems*,

- volume 36, pages 53728–53741. Curran Associates, Inc.
- Sara Rajaei, Sebastian Vincent, Alexandre Berard, Marzieh Fadaee, Kelly Marchisio, and Tom Kocmi. 2026. [Unlocking reasoning capability on machine translation in large language models](#). *Preprint*, arXiv:2602.14763.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. [CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. [GPQA: A Graduate-Level Google-Proof Q&A Benchmark](#). In *First Conference on Language Modeling*, Philadelphia, Pennsylvania, USA.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models](#). *Preprint*, arXiv:2402.03300.
- Ben Shi, Michael Tang, Karthik R Narasimhan, and Shunyu Yao. 2024. [Can Language Models Solve Olympiad Programming?](#) In *First Conference on Language Modeling*, Philadelphia, Pennsylvania, USA.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. [Sequence to Sequence Learning with Neural Networks](#). In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. [Challenging BIG-bench tasks and whether chain-of-thought can solve them](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open Foundation and Fine-Tuned Chat Models](#). *Preprint*, arXiv:2307.09288.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. [Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.
- Karthik Valmeekam, Vardhan Palod, Kaya Stechly, Atharva Gundawar, and Subbarao Kambhampati. 2026. [Beyond semantics: The unreasonable effectiveness of reasonless intermediate tokens](#). *Transactions on Machine Learning Research*. J2C Certification.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is All you Need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Galouédec. 2020. TRL: Transformer Reinforcement Learning. <https://github.com/huggingface/trl>.
- Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025a. [DRT: Deep reasoning translation via long chain-of-thought](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6770–6782, Vienna, Austria. Association for Computational Linguistics.
- Jiaan Wang, Fandong Meng, and Jie Zhou. 2025b. [Ex-trans: Multilingual deep reasoning translation via exemplar-enhanced reinforcement learning](#). *arXiv preprint arXiv:2505.12996*.
- Jiaan Wang, Fandong Meng, and Jie Zhou. 2026. [Deep-Trans: Deep reasoning translation via reinforcement learning](#). *Transactions of the Association for Computational Linguistics*, 14:47–63.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. [Self-Consistency Improves Chain of Thought Reasoning in Language Models](#). In *The Eleventh International Conference on Learning Representations*, Kigali, Rwanda.
- Yifan Wang, Zewei Sun, Shanbo Cheng, Weiguo Zheng, and Mingxuan Wang. 2023b. [Controlling styles in neural machine translation with activation prompt](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2606–2620, Toronto, Canada. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz,

- Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [HuggingFace’s Transformers: State-of-the-art Natural Language Processing](#). *Preprint*, arXiv:1910.03771.
- Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. [A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models](#). In *The Twelfth International Conference on Learning Representations*, Vienna, Austria.
- Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024b. [LLMRefine: Pinpointing and refining large language models via fine-grained actionable feedback](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1429–1445, Mexico City, Mexico. Association for Computational Linguistics.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of Thoughts: Deliberate Problem Solving with Large Language Models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.
- Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, and Denny Zhou. 2024. [Large language models as analogical reasoners](#). In *The Twelfth International Conference on Learning Representations*, Vienna, Austria.
- Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. 2024. [MAMmoTH: Building math generalist models through hybrid instruction tuning](#). In *The Twelfth International Conference on Learning Representations*, Vienna, Austria.
- Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2024. [Tree of problems: Improving structured problem solving with compositionality](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18028–18047, Miami, Florida, USA. Association for Computational Linguistics.
- Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025a. [Compositional translation: A novel LLM-based approach for low-resource machine translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22328–22357, Suzhou, China. Association for Computational Linguistics.
- Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025b. [In-context example selection via similarity search improves low-resource machine translation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.
- Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025c. [TopXGen: Topic-diverse parallel data generation for low-resource machine translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22358–22381, Suzhou, China. Association for Computational Linguistics.
- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. [STaR: Bootstrapping Reasoning With Reasoning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 15476–15488. Curran Associates, Inc.
- Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. [Automatic Chain of Thought Prompting in Large Language Models](#). In *The Eleventh International Conference on Learning Representations*, Kigali, Rwanda.
- Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. [Hunyuan-MT Technical Report](#). *Preprint*, arXiv:2509.05209.
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. [Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Reproducibility Details

A.1 Models, Datasets and Tools

In Table 1, we list the links to the relevant resources used for the experiments.

A.2 Implementation Details

We use HuggingFace’s Transformers library (Wolf et al., 2020; Gugger et al., 2022). We adopt the prompt template introduced by Xu et al. (2024a), and compute the loss only on the target

<i>Datasets</i>	
FLORES-200	https://huggingface.co/datasets/facebook/flores
NTREX HF	https://huggingface.co/datasets/mteb/NTREX
TICO-19	https://huggingface.co/datasets/gmnlp/tico19
<i>Models evaluated</i>	
Qwen3-0.6B	https://huggingface.co/Qwen/Qwen3-0.6B
Qwen3-1.7B	https://huggingface.co/Qwen/Qwen3-1.7B
Qwen3-4B	https://huggingface.co/Qwen/Qwen3-4B
Qwen3-8B	https://huggingface.co/Qwen/Qwen3-8B
Qwen3-14B	https://huggingface.co/Qwen/Qwen3-14B
Qwen3-32B	https://huggingface.co/Qwen/Qwen3-32B
DeepSeek-R1-Distill-Qwen-1.5B	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B
DeepSeek-R1-Distill-Qwen-7B	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
DeepSeek-R1-Distill-Llama-8B	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B
DeepSeek-R1-Distill-Qwen-14B	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B
DeepSeek-R1-Distill-Qwen-32B	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-32B
DeepSeek-R1-Distill-Llama-70B	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-70B
Gemma-3-27B-It	https://huggingface.co/google/gemma-3-27b-it
Gemma-3-4B-Pt	https://huggingface.co/google/gemma-3-4b-pt
Gemma-3-1B-Pt	https://huggingface.co/google/gemma-3-1b-pt
Llama-4-Scout-17B-16E-Instruct	https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct
<i>Other resources</i>	
MetricX24-Hybrid-XXL	https://huggingface.co/google/metricx-24-hybrid-xxl-v2p6
XCOMET-XXL	https://huggingface.co/Unbabel/XCOMET-XXL
FastText	https://huggingface.co/facebook/fasttext-language-identification
vLLM (Kwon et al., 2023)	https://github.com/vllm-project/vllm

Table 1: Links to datasets, benchmarks and models.

(translation or intermediate tokens followed by the translation). We use the same prompt when evaluating the checkpoints. During CoTFT the target is formatted as `<think>\n{Intermediate Tokens}\n</think>\n\nFinal Translation\n{Target translation}`.

Translate this from English to Hausa:
English: "We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.
Hausa:

For instruction-following models and reasoning models we use the following evaluation prompt.

Please write a high-quality Xhosa translation of the following English sentence

"We now have 4-month-old mice that are non-diabetic that used to be diabetic," he added.

Please provide only the translation, nothing more.

We calculate statistical significance using bootstrap resampling (Koehn, 2004) with 300 samples of 500 sentences and a p -value threshold of 0.05.

A.3 CoT Construction Templates

We use the six CoT construction templates proposed by He et al. (2025). They mimic reasoning strategies for translation commonly adopted by human translators.

T1 is Hierarchical Translation

```
<think>
1. Analyze the sentence structure and identify the core elements (subject, verb, object).
2. Translate the sentence from the origin language to the target language, focusing on the core elements.
3. Review the translation for basic accuracy and grammatical structure.
4. Identify areas that need further refinement (e.g., word choice, tense, or word order).
5. Modify the translation to improve fluency and coherence, considering the context.
6. Finalize the translation by ensuring it retains the original meaning while improving readability.
</think>
```

T2 is Triangulating Translation

```
<think>
1. Identify basic elements: Break down the sentence into its main components and identify the key subject, verb, and object.
```

2. Translate to intermediate language: Convert these elements into an intermediate language structure (e.g., simple syntactic rules or function names).
3. Refine back to target language: Translate from the intermediate language back to the target language, adjusting for syntactic norms and idiomatic expressions.
4. Check for accuracy: Ensure that the meaning is preserved in the translated sentence by checking noun-verb agreement and connectors.
5. Adjust word order: Modify word order to ensure that it aligns with the target language's grammatical structure.
6. Final refinement: Review the translation for naturalness, idiomatic use, and overall flow.

T3 is **Back Translation**

- <think>
1. Analyze the provided context in the source language.
 2. Translate the source text to the target language.
 3. Perform back translation from the target language to the source language.
 4. Compare the back translation with the original source context.
 5. Evaluate whether the meaning of the back translation aligns with the original.
 6. If discrepancies are identified, adjust the target language translation to enhance consistency with the original meaning.
 7. Finalize the translation by ensuring both forward and back translations accurately align across all languages involved.
- </think>

T4 is **Context-aware Translation**

- <think>
1. Analyze the current sentence, along with the previous sentences, to understand the overall conversation context.
 2. Identify key elements like tone, formality, or subject matter based on the ongoing conversation.
 3. Translate the sentence while ensuring that the translation is aligned with the tone, style, and subject of the preceding dialogue.
 4. If any ambiguity exists in the translation due to context, refine the translation to better fit the conversation flow.
 5. Verify that the translation maintains coherence with the larger conversation, ensuring consistency in language and tone.

6. Finalize the translation by cross-checking it with the conversation's context to ensure it feels natural and appropriately aligned.
- </think>

T5 is **Translation Explanation**

- <think>
1. Analyze the source sentence and identify the key elements (verbs, subjects, objects, etc.).
 2. Based on these elements, determine the most suitable translation strategy (literal vs. idiomatic).
 3. Select the best translation for each word or phrase, considering context and languagespecific structures.
 4. Explain the rationale behind choosing specific words or phrases.
 5. After completing the initial translation, review each translation decision and explain any adjustments made for fluency or accuracy.
 6. Provide a final explanation for the translation choices, discussing any trade-offs made between literal meaning and contextual appropriateness.
- </think>

T6 is **Structural Transformation**

- <think>
1. Analyze the sentence's syntactic structure in the source language (e.g., identify whether it's active or passive).
 2. Determine the most appropriate syntactic structure in the target language (e.g., whether it needs to be rephrased from active to passive or vice versa).
 3. Adjust the word order and grammatical structure in the target language to match the sentence's meaning, while maintaining clarity.
 4. Translate the sentence, ensuring that subject-verb-object relationships and other syntactic elements align with target language norms.
 5. After the translation, check the sentence's grammar and overall flow in the target language, making sure it is clear and fluid.
 6. If the sentence feels awkward or unnatural, refine the structure by adjusting word choice or reordering components.
- </think>

These approaches serve as more structured alternatives to the basic T0 strategy, where the teacher is simply prompted to choose whatever translation procedure it deems most appropriate for the input.

Explain step by step how to translate the source sentence into the target sentence.

Given a CoT template, we use the following prompt to obtain a CoT produced by a teacher explaining how to obtain the provided translation of a given source sentence following the strategy corresponding to the template. The CoT produced is in the first-person and can later be used for COTFT.

```
Assume that you are a student engaged in
  translating a sentence from {src}
  to {tgt}.
Now you have both the source sentence
  and the target sentence, and need to
  analyze how to translate
  from the source sentence to the given
  target sentence based on the
  provided Thinking Chain Guide. And
  output the chain-of-thought trajectory
  from source to target sentence.

The {src} statement is as follows:
<Source Sentence>
{sentence}
</Source Sentence>

The {tgt} statement is as follows:
<Target Sentence>
{translation}
</Target Sentence>

You continuously reflect on how to
  translate the source sentence to the
  given target sentence
  based on the thinking guidance provided.

The given Thinking Chain Guide is as
  follows:
<Thinking Chain Guide>
{chain_of_thought_template}
</Thinking Chain Guide>

Please refine the entire analysis
  process into a complete self-
  reflective description (in the
  present tense). For self-reflection,
  you can refer to the following
  thinking steps:
directly output the self-reflective
  description in the <think></think>
  tags, without any additional
  descriptions or explanations.
Each line in the reflective description
  can be viewed as a reasoning step in
  the translation process.
```

A.4 Prompting strategies

Step-by-Step Translation (SBYS):

```
{predrafting research}

{draft translation}

Now let's move to the next stage: Post-
  editing with local refinement.
In this stage, the primary aim is to
  refine the draft translation by
  making micro-level improvements that
  improve the draft's fluency.
```

```
Here is a refined version of the
  translation
{refinement}
```

```
Now, we will proofread the refined text
  for grammar spelling, punctuation,
  terminology and overall fluency."
```

```
Here is the translation after
  proofreading
{proofreading}
```

```
We will further improve it to obtain the
  final, polished translation.
```

Multi-Aspect Prompting and Selection (MAPS):

```
Here is a draft translation

1. {zero-shot translation}

Let's write an English sentence related
  to but different from the input
  English sentence and translate it
  into {language}

{demonstrations}

Given this knowledge, we can draft
  another translation

2. {demonstrations-inspired translation}

Let's extract the keywords in the
  provided English sentence, and then
  translate these keywords into {
  language}

{keywords}

Given this knowledge, we can draft
  another translation

3. {keywords-inspired translation}

Let's use a few words to describe the
  topics of the provided English
  sentence

{topics}

Given this knowledge, we can draft
  another translation

4. {topics-inspired translation}

We will choose the best of these
  translations and further improve it
  to obtain the final, polished
  translation.
```

Self-Refine

```
Here is a draft translation

1. {draft translation}
```

```

Let's improve it and write a better
translation

2. {refinement 1}

Let's further improve it and write a
better translation

3. {refinement 2}

Let's improve it one last time and write
a better translation

4. {refinement 3}

We will choose the best of these
translations and further improve it
to obtain the final, polished
translation.

```

Translate, Estimate and Refine (TEaR)

```

Here is a draft translation

1. {draft translation}

Let's identify errors and assess the
quality of the draft translation.
The categories of errors are accuracy (
addition, mistranslation, omission,
untranslated text), fluency (
character encoding, grammar,
inconsistency, punctuation, register
, spelling), locale convention (
currency, date, name, telephone, or
time format) style (awkward),
terminology (inappropriate for
context, inconsistent use), non
translation, other, or no-error.
Each error is classified as one of three
categories: critical, major, and
minor. Critical errors inhibit
comprehension of the text. Major
errors disrupt the flow, but what
the text is trying to say is still
understandable. Minor errors are
technical errors but do not disrupt
the flow or hinder comprehension.

Here are the MQM annotations of the
draft:
{MQM annotations}

Upon reviewing the translation and error
information, we can refine the
draft and obtain a better
translation

2. {refinement}

We will further improve it to obtain the
final, polished translation."

```

Compositional Translation (CompTra)

```

1. English Sentence
{}
Xhosa Translation
{}

```

```

2. English Sentence
{}
Xhosa Translation
{}

3. English Sentence
{}
Xhosa Translation
{}

```

A.5 RL Training Hyperparameters

For GRPO, we use the Hugging Face TRL library (von Werra et al., 2020). Training is conducted on four H100 GPUs, with one dedicated to model deployment for reward computation. We set a per-device batch size of 4 with 4 gradient accumulation steps, for a total of 5000 steps including 100 warmup steps. Hyperparameters include a beta value of 0.02, a maximum gradient norm of 1.0, and a temperature of 1.0. For generation, we sample 12 outputs per prompt with an effective batch size of 48. We apply LoRA (Hu et al., 2022), fine-tuning the q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, and down_proj modules with rank $r = 32$, scaling factor $\alpha = 64$, and dropout rate 0.05.

B Additional Experiments

B.1 Benchmarking LRMs at Scale: To Think, or Not To Think?

In this section we further investigate the impact of thinking tokens when benchmarking LRMs. In Table 3 we report the chrF++⁵ (Popović, 2015; Popović, 2017) and XCOMET-XXL (Guerreiro et al., 2024) scores in the same setup as BLEU and MetricX in Table 2. They tell the same story as BLEU and MetricX. Outputting thinking tokens only marginally helps; the gains are not consistent and when they occur they are small. This questions the necessity of an LRM to think before doing MT, all the more so that thinking is considerably more expensive than straight up answering. It is worth noting that “small” models (Qwen-0.6B, Qwen-1.7B and Qwen-4B) often generate answers in English or Chinese when they struggle with the target language (e.g., Czech, Finnish, Kazakh, Lithuanian etc.) resulting in artificially better neural scores. The thinking mode particularly helps in such scenarios because it allows the model to remember that it should write an answer in a different language

⁵nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.4.2

Models	Czech		Finnish		French		German		Japanese	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
QWEN3-0.6B	4.63	22.86	2.23	23.64	24.65	9.03	16.00	9.95	7.65	8.67
<i>w/o Thinking</i>	3.99	13.60	2.80	21.16	23.33	8.52	15.27	8.21	5.71	7.56
QWEN3-1.7B	14.73	15.58	5.83	20.29	37.98	4.80	27.86	4.50	15.49	5.84
<i>w/o Thinking</i>	15.51	15.09	7.08	19.38	38.08	4.44	28.03	3.97	15.08	5.87
QWEN3-4B	23.82	9.05	13.47	14.22	45.40	3.19	35.06	2.58	19.87	5.06
<i>w/o Thinking</i>	24.43	9.27	13.59	14.30	44.69	3.14	34.64	2.48	20.42	4.82
QWEN3-8B	30.11	6.38	19.29	9.95	48.72	2.59	39.29	1.62	23.45	4.31
<i>w/o Thinking</i>	30.27	6.61	19.21	10.20	49.03	2.49	39.05	1.56	24.43	4.08
QWEN3-14B	34.07	4.99	22.73	7.74	51.21	2.26	42.39	1.29	26.25	3.80
<i>w/o Thinking</i>	33.55	5.16	23.17	7.67	51.88	2.12	41.64	1.16	27.31	3.75
QWEN3-32B	34.62	4.70	24.33	7.25	51.80	2.01	42.78	1.09	26.44	3.88
<i>w/o Thinking</i>	33.27	4.64	24.14	7.05	50.94	2.00	42.08	1.07	27.33	3.62

Models	Kazakh		Lithuanian		Portuguese		Spanish		Turkish	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
QWEN3-0.6B	0.41	23.61	1.28	24.50	25.58	9.11	17.80	7.67	5.88	21.66
<i>w/o Thinking</i>	0.49	22.26	1.77	23.78	23.31	8.68	17.00	7.04	6.27	20.00
QWEN3-1.7B	0.92	23.41	5.03	21.70	39.02	4.46	25.43	3.87	13.20	15.48
<i>w/o Thinking</i>	1.36	23.54	5.83	21.56	39.01	4.35	25.45	3.75	14.35	13.94
QWEN3-4B	8.02	16.02	12.83	15.69	46.08	2.97	28.46	2.69	21.49	9.75
<i>w/o Thinking</i>	8.15	16.37	13.05	15.45	45.80	3.09	28.70	2.69	21.74	9.55
QWEN3-8B	13.76	11.68	17.49	11.51	48.93	2.53	30.77	2.15	27.09	7.28
<i>w/o Thinking</i>	12.98	11.76	18.03	11.30	48.76	2.41	30.60	2.03	26.90	6.76
QWEN3-14B	17.81	8.87	22.06	8.78	50.47	2.25	31.77	1.95	30.72	5.93
<i>w/o Thinking</i>	17.42	8.84	22.82	8.45	51.42	2.09	32.02	1.86	29.85	5.70
QWEN3-32B	17.33	9.54	23.70	8.41	51.01	1.96	31.84	1.66	31.68	5.80
<i>w/o Thinking</i>	18.13	8.39	24.10	7.58	51.59	1.96	32.18	1.71	30.48	5.59

Table 2: BLEU and MetricX scores for ten English $\rightarrow$ X directions from FLORES 200, with thinking (first line) and without thinking (second line). Best results are highlighted in bold (higher for BLEU and lower for MetricX).

than what it is “used” to generating in. When the models are big enough (typically $\geq 8B$), thereby solving this incorrect language issue, thinking does not result in any gains. Moreover, we run additional experiments with more DeepSeek-R1-Distill models (see Table 4) and again we observe a similar pattern, “no thinking” consistently outperforms thinking. We run additional experiments when translating in the 5-shot setting, retrieving demonstrations from the FLORES 200 dev test with bm25s (Lù, 2024) following Zebaze et al. (2025b). As shown in Table 5, providing demonstrations does not help the thinking mode take performance to an upper level. Similarly, it does not help when translating into English, as reported in Table 6.

B.2 Human Evaluation

We carried out a focused human comparison of “thinking” versus “not thinking” for English $\rightarrow$ French translation using Qwen3-32B. We randomly sampled 100 FLORES-200 examples and carried out reference-free pairwise ranking (better, worse, same) with native French speakers. To avoid

biasing the results against the reasoning model, we excluded instances where it failed to produce a translation due to getting lost in its CoT. Annotators judged both translations as equally good or bad in 36% of cases; the “thinking” answers were preferred in 31% of cases, and the “not thinking” answers in 33%. We conducted the same experiment in a reference-based fashion to evaluate Turkish $\rightarrow$ English and found that both translations are equally good or bad in 47% of the cases; “thinking” answers were preferred in 30% of cases, and the “not thinking” answers in 23%. Overall, these findings corroborate our conclusions in Section 3 that MT capabilities of RM do not significantly benefit from thinking tokens.

B.3 Results on NTREX-128 and TICO-19

In this section, we evaluate the models on 2 additional benchmarks:

- **NTREX 128** (Barrault et al., 2019; Federmann et al., 2022) is an MT benchmark derived from WMT19 news data translated by professional human translators. It contains 1997 parallel sen-

Models	Czech		Finnish		French		German		Japanese	
	chrF++	XCOMET	chrF++	XCOMET	chrF++	XCOMET	chrF++	XCOMET	chrF++	XCOMET
QWEN3-0.6B	19.43	16.52	16.36	13.56	47.19	53.77	39.34	71.05	14.11	44.66
<i>w/o Thinking</i>	18.53	55.22	18.12	23.62	45.94	54.82	37.75	74.96	11.35	53.73
QWEN3-1.7B	33.89	39.91	26.45	20.00	57.40	80.58	49.52	88.42	21.39	69.82
<i>w/o Thinking</i>	35.27	38.62	29.54	22.68	57.68	80.67	49.89	89.58	20.81	67.83
QWEN3-4B	43.43	68.22	36.99	45.99	62.49	89.14	55.13	93.48	24.68	80.26
<i>w/o Thinking</i>	43.92	65.45	38.11	42.54	62.26	88.23	55.07	94.25	26.53	79.23
QWEN3-8B	48.49	80.32	43.15	65.78	64.68	92.19	58.30	95.94	27.24	86.18
<i>w/o Thinking</i>	48.70	77.64	43.44	62.60	65.16	91.89	58.21	96.38	28.50	86.60
QWEN3-14B	51.57	86.91	46.26	76.27	66.41	92.99	60.52	96.89	29.88	89.88
<i>w/o Thinking</i>	51.29	84.65	46.78	74.75	66.91	93.57	60.18	97.37	30.54	89.76
QWEN3-32B	51.45	86.59	46.47	77.18	65.99	92.79	60.35	96.27	29.10	89.38
<i>w/o Thinking</i>	50.85	86.48	47.40	77.13	65.98	93.81	60.38	97.60	30.89	89.97

Models	Kazakh		Lithuanian		Portuguese		Spanish		Turkish	
	chrF++	XCOMET	chrF++	XCOMET	chrF++	XCOMET	chrF++	XCOMET	chrF++	XCOMET
QWEN3-0.6B	4.14	15.77	10.73	14.19	48.05	67.21	41.70	70.45	22.64	18.39
<i>w/o Thinking</i>	6.78	20.31	12.97	16.91	44.73	68.36	40.29	72.40	23.77	22.97
QWEN3-1.7B	7.51	15.09	24.83	18.80	58.97	87.43	48.25	87.84	34.08	39.47
<i>w/o Thinking</i>	10.84	15.62	26.11	19.52	58.99	87.07	48.61	87.55	35.89	43.02
QWEN3-4B	26.54	25.20	36.26	43.79	63.93	92.69	50.66	92.23	42.97	66.46
<i>w/o Thinking</i>	26.51	23.78	36.60	42.88	63.70	92.17	51.07	92.20	43.52	65.43
QWEN3-8B	34.64	40.48	41.54	61.88	65.66	94.16	52.45	94.13	47.86	77.71
<i>w/o Thinking</i>	33.86	36.52	42.02	60.84	65.77	94.60	52.55	94.62	47.90	77.68
QWEN3-14B	39.53	53.44	45.08	73.81	66.70	95.12	53.29	94.64	50.97	83.68
<i>w/o Thinking</i>	38.95	50.77	45.82	73.72	67.55	95.31	53.60	95.25	50.42	83.35
QWEN3-32B	38.34	53.66	45.92	75.60	66.71	94.93	53.19	95.07	51.29	84.03
<i>w/o Thinking</i>	39.22	52.05	46.68	76.89	67.64	95.77	53.57	95.32	50.60	83.08

Table 3: chrF++ and XCOMET scores for 10 English → X directions from FLORES 200. Best results are highlighted in bold.

tences and is recommended for the evaluation of from-English translation directions. We use the first 1000 sentence pairs for evaluation, and the last 997 sentence pairs as the selection pool.

- **TICO-19** (Anastasopoulos et al., 2020) is an MT benchmark comprising texts on the COVID-19 pandemic covering 35 languages. Its validation and test sets consist of 971 (used as a selection pool) and 2100 samples respectively.

We focus on translating from English. We report the results obtained on NTREX 128 in Table 7 and those obtained on TICO-19 in Table 8. We reach the same conclusions on NTREX 128 as with FLORES 200. On TICO-19 we consider different languages, mostly from Asia. Despite some of them being low-resource languages (Khmer, Marathi etc.), thinking offers little to no advantage.

Models	Czech		Finnish		French		German		Japanese	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
DEEPSEEK-R1-DISTILL-QWEN-14B	22.78	9.91	10.55	17.40	46.13	2.87	34.54	2.49	21.04	4.85
<i>w/o Thinking</i>	23.04	9.50	10.00	17.13	45.42	2.54	30.26	2.40	22.04	4.45
DEEPSEEK-R1-DISTILL-QWEN-32B	28.49	7.12	15.61	13.06	48.68	2.78	38.25	1.82	24.11	4.56
<i>w/o Thinking</i>	29.61	6.22	16.12	12.52	49.35	2.16	38.95	1.44	26.57	3.89
DEEPSEEK-R1-DISTILL-LLAMA-70B	37.31	4.19	29.90	5.35	52.34	2.22	43.39	1.16	25.61	4.08
<i>w/o Thinking</i>	38.47	3.52	30.96	4.38	52.34	1.88	44.78	0.91	27.94	3.62
Models	Kazakh		Lithuanian		Portuguese		Spanish		Turkish	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
DEEPSEEK-R1-DISTILL-QWEN-14B	2.38	13.25	7.87	20.02	46.28	2.86	29.48	2.56	17.99	12.79
<i>w/o Thinking</i>	2.50	6.49	7.21	18.79	47.11	2.52	30.52	2.13	17.78	11.70
DEEPSEEK-R1-DISTILL-QWEN-32B	5.66	18.37	13.24	16.09	49.17	2.55	30.62	2.26	24.52	9.36
<i>w/o Thinking</i>	4.95	16.19	14.02	14.92	50.53	2.02	31.67	1.74	24.96	7.88
DEEPSEEK-R1-DISTILL-LLAMA-70B	21.39	8.14	25.63	8.22	52.27	2.09	32.68	1.78	33.42	5.46
<i>w/o Thinking</i>	21.56	6.76	27.01	7.33	52.64	1.75	33.54	1.55	34.77	5.00

Table 4: BLEU and MetricX scores for 10 English $\rightarrow$ X directions from FLORES 200. Best results are highlighted in bold.

Models	Czech		Finnish		French		German		Japanese	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
QWEN3-0.6B	5.65	22.53	3.21	23.13	24.49	9.51	16.11	10.15	8.08	9.29
<i>w/o Thinking</i>	5.31	22.49	2.69	23.40	21.48	10.28	14.51	10.43	6.09	11.84
QWEN3-1.7B	15.92	14.68	7.86	19.18	38.28	4.69	28.41	4.30	16.46	5.85
<i>w/o Thinking</i>	16.26	14.48	8.50	18.71	38.52	4.54	28.19	4.05	15.29	5.89
QWEN3-4B	24.43	9.18	13.96	13.88	44.47	3.66	34.20	3.14	20.68	5.22
<i>w/o Thinking</i>	25.77	8.74	15.33	13.39	44.99	3.25	34.79	2.54	21.36	4.79
QWEN3-8B	30.11	6.65	19.37	10.21	48.89	2.89	38.88	1.92	24.88	4.28
<i>w/o Thinking</i>	30.36	6.59	19.70	9.89	49.18	2.58	39.16	1.59	24.67	4.10
QWEN3-14B	34.44	5.11	23.51	7.59	51.23	2.37	41.83	1.39	27.93	3.86
<i>w/o Thinking</i>	34.01	4.93	23.89	7.43	51.45	2.13	41.76	1.17	28.00	3.63
QWEN3-32B	15.69	15.01	11.19	16.04	29.14	11.67	22.99	11.00	16.14	12.00
<i>w/o Thinking</i>	34.34	4.64	24.91	6.72	50.35	1.99	42.58	1.09	27.86	3.53
Models	Kazakh		Lithuanian		Portuguese		Spanish		Turkish	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
QWEN3-0.6B	0.93	23.89	2.21	24.04	26.00	9.44	17.51	8.38	6.46	21.24
<i>w/o Thinking</i>	1.31	23.85	2.11	24.12	21.99	11.08	15.81	9.11	5.43	21.16
QWEN3-1.7B	2.03	22.03	6.80	20.51	39.27	4.41	25.76	3.85	14.10	14.36
<i>w/o Thinking</i>	3.24	21.58	7.69	20.65	39.22	4.23	25.63	3.76	14.71	13.41
QWEN3-4B	10.1	15.26	13.64	15.10	45.61	3.44	28.43	3.14	21.32	10.28
<i>w/o Thinking</i>	11.0	14.50	14.62	14.44	45.93	3.16	29.12	2.69	21.95	9.63
QWEN3-8B	15.64	11.07	19.01	10.97	49.03	2.64	30.87	2.30	27.01	7.34
<i>w/o Thinking</i>	15.77	10.80	19.51	10.67	49.44	2.47	31.19	2.09	26.98	6.82
QWEN3-14B	20.35	8.30	23.68	8.41	50.93	2.28	32.38	1.94	30.99	6.08
<i>w/o Thinking</i>	19.40	7.97	24.59	7.86	51.37	2.10	32.67	1.72	29.97	5.66
QWEN3-32B	10.85	15.88	11.89	16.26	29.30	11.61	17.69	11.53	16.08	14.69
<i>w/o Thinking</i>	20.63	7.61	25.28	7.12	50.89	1.95	33.05	1.58	30.97	5.26

Table 5: 5-shot BLEU and MetricX scores for 10 English $\rightarrow$ X directions from FLORES 200. Best results are highlighted in bold.

Models	Czech		Finnish		French		German		Japanese	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
QWEN3-0.6B	23.33	7.29	11.40	13.01	35.95	3.32	33.10	4.12	15.47	5.96
<i>w/o Thinking</i>	22.05	7.42	9.30	11.40	35.02	3.31	31.26	4.24	14.88	6.38
QWEN3-1.7B	34.90	3.33	23.42	6.25	41.40	1.99	40.50	2.29	23.67	3.03
<i>w/o Thinking</i>	34.23	3.54	22.62	6.25	41.89	2.03	40.94	2.31	22.66	3.41
QWEN3-4B	38.76	2.15	30.57	3.68	44.54	1.59	44.43	1.59	27.05	2.26
<i>w/o Thinking</i>	39.06	2.26	30.70	3.47	45.59	1.61	44.83	1.66	27.32	2.40
QWEN3-8B	40.02	1.80	33.28	2.52	45.80	1.42	45.14	1.50	27.98	1.92
<i>w/o Thinking</i>	40.77	1.86	33.49	2.46	46.54	1.43	45.46	1.42	28.95	1.93
QWEN3-14B	41.91	1.60	35.45	2.12	46.70	1.37	46.28	1.36	29.20	1.75
<i>w/o Thinking</i>	43.28	1.57	36.31	2.00	48.64	1.32	47.49	1.32	30.43	1.77
QWEN3-32B	43.19	1.47	37.01	1.84	47.51	1.27	46.77	1.29	29.99	1.68
<i>w/o Thinking</i>	44.23	1.42	37.84	1.77	48.72	1.26	47.52	1.26	30.88	1.70
Models	Kazakh		Lithuanian		Portuguese		Spanish		Turkish	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
QWEN3-0.6B	6.94	15.07	10.21	13.71	39.22	3.59	25.92	3.92	16.12	9.21
<i>w/o Thinking</i>	5.92	15.67	4.36	7.41	37.72	3.70	25.64	3.99	13.83	9.34
QWEN3-1.7B	17.90	8.48	23.61	6.26	45.63	2.15	29.79	2.43	29.31	4.39
<i>w/o Thinking</i>	16.20	9.17	22.23	6.54	46.32	2.31	30.49	2.46	27.65	4.53
QWEN3-4B	25.64	5.02	29.64	3.75	49.12	1.79	32.59	1.95	34.96	2.77
<i>w/o Thinking</i>	24.20	5.31	29.37	3.83	50.16	1.84	33.55	2.02	34.69	2.88
QWEN3-8B	29.31	3.80	33.01	2.76	50.04	1.54	32.89	1.75	37.88	2.15
<i>w/o Thinking</i>	28.91	3.98	33.28	2.81	50.89	1.56	33.80	1.78	37.77	2.24
QWEN3-14B	32.01	3.06	34.31	2.48	51.22	1.48	34.06	1.63	39.52	1.88
<i>w/o Thinking</i>	32.17	3.08	35.15	2.43	53.28	1.49	35.50	1.63	40.18	1.86
QWEN3-32B	33.33	3.27	36.10	2.06	51.69	1.44	34.49	1.57	40.91	1.81
<i>w/o Thinking</i>	34.03	2.77	37.15	2.14	53.66	1.39	35.85	1.54	41.87	1.72

Table 6: BLEU and MetricX scores for 10 X → English directions from FLORES 200. Best results are highlighted in bold.

Models	Czech		Finnish		French		German		Japanese	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
QWEN3-0.6B	4.62	22.82	3.03	23.26	16.80	10.38	13.93	10.38	5.66	10.06
<i>w/o Thinking</i>	4.92	15.63	3.33	21.09	16.72	9.88	13.89	9.18	4.18	8.78
QWEN3-1.7B	13.64	16.44	6.93	19.37	26.14	5.86	24.33	4.97	12.93	6.94
<i>w/o Thinking</i>	14.11	16.13	6.95	19.19	25.61	6.13	23.43	4.98	12.41	7.03
QWEN3-4B	22.23	9.92	12.70	14.04	31.07	4.09	30.41	2.97	18.11	5.64
<i>w/o Thinking</i>	22.06	10.16	11.83	14.56	30.82	4.43	30.21	3.16	16.77	5.91
QWEN3-8B	28.40	7.19	17.02	10.28	33.10	3.51	34.19	1.98	19.55	5.09
<i>w/o Thinking</i>	27.54	7.92	16.63	10.82	33.44	3.51	34.27	2.20	19.99	5.15
QWEN3-14B	31.04	5.68	20.22	7.84	34.61	3.07	36.85	1.68	22.02	4.65
<i>w/o Thinking</i>	29.92	6.18	19.86	8.25	34.80	3.17	36.82	1.69	21.88	4.61
QWEN3-32B	32.63	5.73	21.32	7.47	34.76	2.95	37.09	1.53	21.19	4.97
<i>w/o Thinking</i>	30.48	5.73	21.71	7.38	34.55	2.84	37.13	1.46	20.59	4.51
Models	Kazakh		Lithuanian		Portuguese		Spanish		Turkish	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
QWEN3-0.6B	0.26	23.60	0.90	24.26	19.57	10.00	22.01	8.77	6.76	21.87
<i>w/o Thinking</i>	0.66	22.05	1.46	23.90	18.37	9.28	20.12	8.21	7.04	20.30
QWEN3-1.7B	0.52	23.71	4.62	21.42	29.23	5.49	30.94	4.83	12.58	16.33
<i>w/o Thinking</i>	1.13	23.68	5.06	21.47	28.92	5.60	30.81	4.96	12.73	15.25
QWEN3-4B	5.95	17.29	10.59	15.76	34.60	3.76	36.44	3.32	19.26	11.29
<i>w/o Thinking</i>	5.54	17.81	10.86	15.83	33.83	4.23	35.91	3.62	19.05	11.31
QWEN3-8B	9.70	13.39	14.80	11.97	36.71	3.10	38.91	2.78	23.07	8.67
<i>w/o Thinking</i>	8.93	14.07	14.96	11.92	36.89	3.35	38.85	2.95	23.29	8.84
QWEN3-14B	13.18	10.57	18.89	9.58	38.49	2.83	40.24	2.53	26.35	7.38
<i>w/o Thinking</i>	12.65	11.04	19.42	9.16	38.39	2.91	40.18	2.59	26.31	7.36
QWEN3-32B	13.53	10.93	19.93	9.01	39.21	2.74	40.70	2.41	27.62	6.93
<i>w/o Thinking</i>	13.05	10.40	20.77	8.62	39.23	2.76	41.30	2.35	26.68	6.97

Table 7: BLEU and MetricX scores for 10 English $\rightarrow$ X directions from NTREX 128. Best results are highlighted in bold.

Models	Bengali		Farsi		Hindi		Indonesian		Khmer	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
QWEN3-0.6B	0.07	22.31	1.86	20.03	0.32	21.64	22.66	7.43	0.63	22.87
<i>w/o Thinking</i>	0.32	21.61	3.17	3.46	0.58	21.79	21.18	6.51	1.02	23.80
QWEN3-1.7B	2.85	15.49	10.07	13.50	6.32	14.70	37.04	3.47	1.04	22.10
<i>w/o Thinking</i>	4.69	13.13	10.46	13.56	9.03	13.15	38.33	3.45	1.54	22.21
QWEN3-4B	14.50	6.31	19.11	7.62	24.40	6.23	45.14	2.82	9.20	15.05
<i>w/o Thinking</i>	13.90	6.16	18.59	7.72	23.40	6.34	46.60	2.52	8.27	15.54
QWEN3-8B	18.97	4.13	25.13	4.70	30.22	4.82	48.83	1.89	16.66	10.65
<i>w/o Thinking</i>	17.98	4.15	24.91	4.58	29.95	4.90	50.19	1.94	16.20	10.63
QWEN3-14B	23.68	3.10	29.01	3.55	35.81	4.12	51.21	1.66	22.09	8.21
<i>w/o Thinking</i>	22.87	3.28	29.33	3.38	35.65	4.26	52.07	1.71	22.30	8.41
QWEN3-32B	17.94	7.76	27.39	4.41	31.15	6.83	51.95	1.62	16.52	11.42
<i>w/o Thinking</i>	24.40	2.89	29.52	3.16	37.37	3.98	52.47	1.59	21.22	7.13
Models	Marathi		Malay		Nepali		Tagalog		Urdu	
	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX	BLEU	MetricX
QWEN3-0.6B	0.03	23.58	13.57	9.11	0.24	22.75	5.66	21.11	0.41	22.99
<i>w/o Thinking</i>	0.06	24.04	10.49	4.98	0.61	21.58	7.72	17.76	0.85	22.18
QWEN3-1.7B	1.27	19.50	22.81	4.71	2.45	17.30	10.18	18.67	2.74	18.42
<i>w/o Thinking</i>	2.49	18.36	23.17	4.54	9.83	10.04	10.65	18.64	3.08	18.26
QWEN3-4B	7.78	9.80	31.93	3.99	9.83	10.04	21.62	11.54	11.37	10.82
<i>w/o Thinking</i>	7.75	10.61	32.87	4.07	8.13	11.42	24.30	10.52	10.55	10.97
QWEN3-8B	12.01	6.92	37.30	2.95	15.25	6.93	27.44	7.60	16.35	7.06
<i>w/o Thinking</i>	11.93	7.03	39.25	3.12	14.42	7.53	30.00	7.18	16.56	6.88
QWEN3-14B	15.05	5.35	41.91	2.63	18.23	5.89	32.38	5.35	20.97	5.06
<i>w/o Thinking</i>	15.11	5.49	42.91	2.67	17.93	6.26	34.29	5.33	21.04	5.24
QWEN3-32B	13.26	8.61	43.30	2.57	17.40	8.31	31.44	6.04	20.21	6.36
<i>w/o Thinking</i>	16.63	4.98	44.12	2.65	20.83	5.45	36.12	4.80	21.87	4.87

Table 8: BLEU and MetricX scores for 10 English $\rightarrow$ X directions from TICO-19. Best results are highlighted in bold.

B.4 Does distilled Chain-of-Thought as intermediate tokens improve performance?

Following Section 5.2, we compare CoTFT with IOFT across all six templates using gemma-3-1b-pt as the student and gemma-3-27b-it as the intermediate teacher. We focus on translating from English to Lithuanian. As shown in Figure 9, CoTFT consistently lags behind IOFT. The gap can be as large as 5 BLEU and 4 MetricX points. Despite T3 being the best template, it is still largely behind IOFT in terms of performance.

B.5 What happens when we use traces from MT prompting strategies as intermediate tokens?

In Figure 10, we observe that CoTFT with reasoning traces based on alternative prompting strategies outperforms IOFT. SBYS is an exception, for which CoTFT is behind IOFT. Across prompting strategies, IOFT-MAX outperforms IOFT and CoTFT with the only exception of CompTra. This is exactly what happened with our experiments with Llama-4-Scout-17B-16E-Instruct and gemma-3-4b-pt for Xhosa.

Finally, as shown in Figure 11, CoTFT-MAX fails to improve over IOFT-MAX, confirming our previous conclusions with Xhosa. In this case, CompTra is not an exception. For all the strategies, CoTFT-MAX and IOFT-MAX are very close in performance, with IOFT-BOA topping them all. This again suggests that reasoning traces do not help, even when they are based on MT prompting strategies whose drafting attempts do not outperform the ground truth in terms of quality. Having target translations of high-quality (IOFT-BOA) has the highest impact, outperforming the standard IOFT by 3 BLEU and 1.3 MetricX with the same number of parallel pairs and the same training recipe.

B.6 Down the rabbit hole of sentence decomposition

Following Section 5.4, we evaluate multiple sentence decomposition approaches and compare CoTFT against standard IOFT and IOFT-EXT. As shown in Figure 12, CoTFT consistently outperforms IOFT across all decomposition strategies. Again, SP and CompTra works better than P and H. Using the generated pairs as additional training

samples (i.e. IOFT-EXT) is particularly helpful with P and SP because they correspond to fully-fledged sentences as explained earlier. CoTFT with H and CoTFT with CompTra outperform the corresponding IOFT-EXT suggesting that short phrases and their translations are relevant intermediate information for CoTFT. CoTFT with CompTra works just as well as IOFT-EXT(P) which is impressive as the latter required multiplying the size of $\mathcal{D}$ by six.

B.7 CoT distillation: What happens when we change the teacher?

In this section, we run the same CoT distillation experiment as in Section 5.2 but we use DeepSeek-R1-Distill-Llama-70B instead as the teacher. As seen in Figure 13, CoTFT behaves similarly to IOFT across templates. The performance of CoTFT is better than what we observed with Llama-4-Scout-17B-16E-Instruct despite LLAMA being better at translating into Xhosa. We attribute this to the “thinking” abilities of DEEPSEEK-R1 which, despite not being good at generating Xhosa, can generate a better explanation compared to LLAMA as to why a hypothesis is an accurate translation of a source. However, in both cases, CoTFT does not improve over IOFT, which is also faster to train in comparison.

B.8 MT traces generated by prompting strategies as intermediate tokens: What happens when we change the teacher?

We run additional information on Xhosa and change the teacher from Llama-4-Scout-17B-16E-Instruct to gemma-3-27b-it. GEMMA’s zero-shot MT performance (on FLORES 200, BLEU = 12.82, MetricX = 7.62) is worse than LLAMA’s (BLEU = 16.90, MetricX = 6.53) and we aim to investigate how this impacts our findings. First of all, CoTFT outperforms IOFT across prompting strategies with the exception of SBYS. IOFT-MAX outperforms IOFT and we observe the same behaviour between IOFT-MAX(COMPTRA) and IOFT as we did with LLAMA. Despite CoTFT with SBYS underperforming IOFT, IOFT-MAX(SBYS) outperforms IOFT, meaning that translation attempts embedded in SBYS-inspired CoT are helpful, but they are drowned out by other useless tokens, which impacts how well CoTFT performs. Ultimately, IOFT-BOA works best; although IOFT-MAX(TEAR) achieves higher BLEU

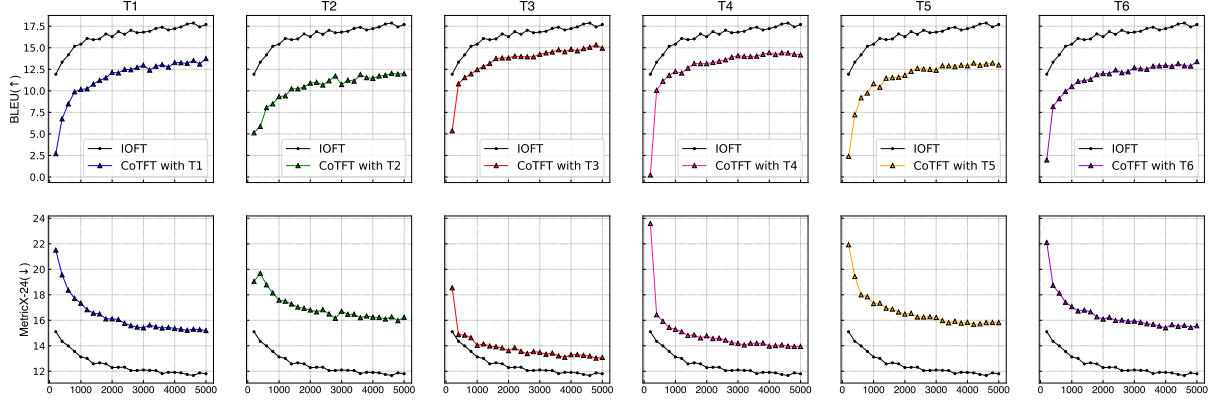


Figure 9: Comparison between CoTFT and IOFT with six different CoT templates.

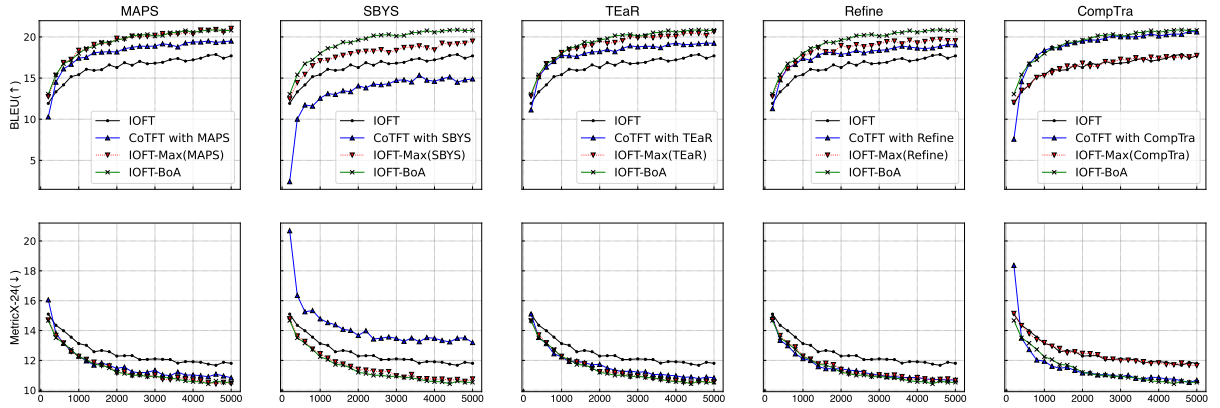


Figure 10: Comparison between IOFT and CoTFT with five different prompting strategies.

scores it lags behind in terms of MetricX. It even outperforms the teacher gemma-3-27b-it despite being six times smaller.

When we use gemma-3-4b-it as the teacher (gemma-3-4b-pt being the student), the traces obtained using prompting strategies do not help CoTFT to outperform IOFT. In Figure 15, we observe a degradation of performance that confirms our intuition suggesting that these traces are helpful only if they contain translation attempts that are better than the ground truth.

B.9 Reinforcement Learning After IOFT and CoFT

We investigate whether CoTFT improves performance during RL fine-tuning. We consider all setups: IOFT, CoTFT with T_3 (like in Section 5.2) and CoTFT with all the modular prompting strategies. The final checkpoints (checkpoint-5000) are further fine-tuned with GRPO (Shao et al., 2024) on a second parallel dataset (on 3 GPUs for 5000 steps, more details in Appendix A.5). We consider three reward functions: one based on the BLEU

and chrF++ scores with the ground truth, a second using COMET-22 (wmt22-cometkiwi-da; Rei et al., 2022) and a last one based on the BLASER-2.0 QE scores between the sources and hypotheses. For CoTFT, we consider an additional format reward to ensure that the models preserve their prior thinking before translating. We compare all RL fine-tunings with IOFT on the second dataset (on 1 GPU for 5000 steps) and report the results in Figure 16. We observe that GRPO reverses the relative performance ordering between IOFT and CoTFT prior to fine-tuning (e.g. CoTFT with $T_3 \leq \text{IOFT} \leq \text{CoTFT with CompTra}$). GRPO results in consistent improvements of approximately +1.3 BLEU and -1.0 MetricX across all setups. Notably, CoTFT still does not outperform IOFT, even with RL. This is consistent with Zheng et al.’s (2025) findings, namely that CoT signals fail to induce meaningful reasoning when the reward is applied only to the final translation. Moreover, unlike mathematics where step-by-step explanations are widely present in pre-training corpora (proofs), it is not the case for translation data. This scarcity of reasoning-

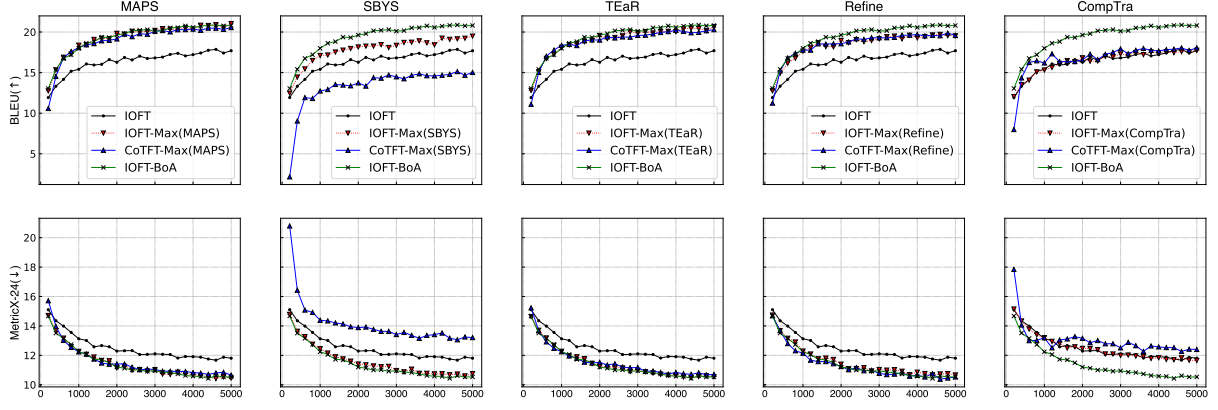


Figure 11: Comparison between IOFT and CoTFT with five different prompting strategies.

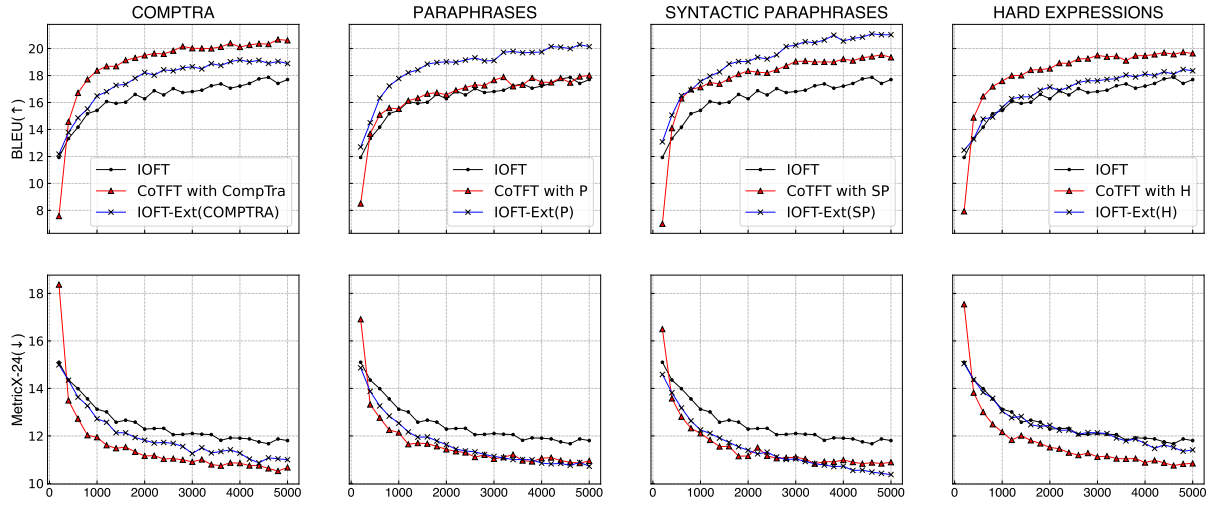


Figure 12: Comparison between IOFT and CoTFT with four different sentence decomposition strategies.

like data may explain CoT’s limited effectiveness in MT. Moreover, in the context of RL for MT, the notion of a “verifiable reward” is not well-defined. The idea of a “correct translation” is far less absolute than a correct answer in mathematics (e.g. $1+1=2$). As a result, we rely on proxy metrics such as BLEU or COMET, but using these rewards do not reproduce the dramatic improvements observed for mathematics and code. Finally, we find that continuing SFT (IOFT) on the IOFT checkpoint gives much larger gains (+6 BLEU, -3 MetricX) than GRPO which quickly stagnates, reinforcing that standard IOFT alone can achieve superior MT performance.

B.10 What happens when we change the student?

We conduct additional English→Xhosa experiments with Llama-4-Scout-17B-16E-Inst as the teacher but replace the student model

gemma-3-4b-pt with gemma-3-12b-pt. To reduce computational costs, we apply LoRA (Hu et al., 2022) while keeping all other hyperparameters unchanged. Overall, CoTFT continues to underperform IOFT across all six CoT distillation templates (Figure 17). The performance gap is noticeably larger than with gemma-3-4b-pt, which we attribute to the use of LoRA. Conversely (Figure 18), traces derived from MT prompting strategies generally serve as better intermediate tokens for CoTFT: CompTra, Self-Refine, and MAPS all surpass IOFT. As before, IOFT-MAX outperforms standard IOFT, and the same pattern holds when comparing IOFT-MAX(COMPTRA) to IOFT, mirroring our observations with the 4B model. Even though CoTFT with SBYS and TEaR underperforms IOFT, their IOFT-MAX variants (IOFT-MAX(SBYS) and IOFT-MAX(TEaR)) achieve higher performance. This suggests that the translation attempts embedded within SBYS and TEaR-

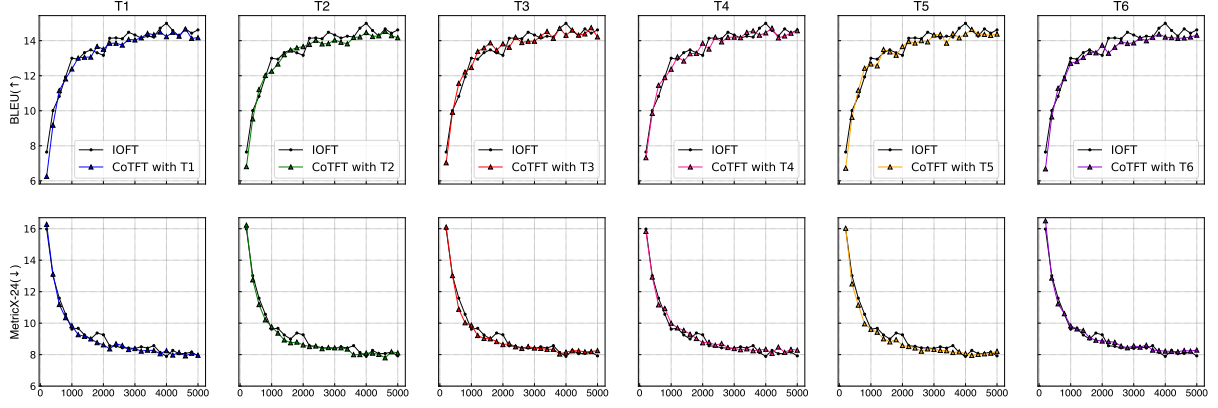


Figure 13: Comparison between IOFT and CoTFT with six different CoT templates.

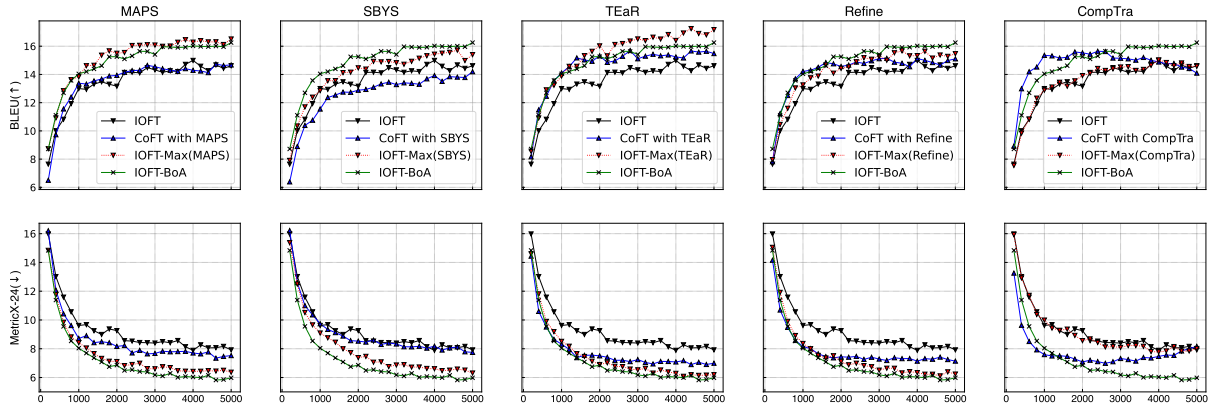


Figure 14: Comparison between IOFT and CoTFT with five different prompting strategies.

style CoT are indeed useful, but their signal is diluted by extraneous tokens when used directly in CoTFT. In line with the rest of our findings, IOFT-BoA achieves the strongest results.

B.11 Insights on Document-level MT and Stylized MT

Following Wang et al. (2023b), we investigate what they refer to as stylized translation: translating into different regional or formal variants of the same language (e.g., Simplified vs. Traditional Chinese, European vs. Brazilian Portuguese). While stylized translation is a research problem in its own right, we include these preliminary experiments to strengthen our overall findings. We note, however, that standard automatic metrics may not fully capture the nuances of style, which remain challenging even for human evaluation. We first study translation between English and three Chinese variants: Simplified, Traditional, and Yue. Experiments are conducted on the FLORES-200 devtest set. We also evaluate English–Portuguese translation, distinguishing between European and Brazilian Por-

tuguese: FLORES-200 is used for European Portuguese, while Brazilian Portuguese is evaluated on TICO-19 (Anastasopoulos et al., 2020). Finally, we consider Korean honorific and non-honorific forms.⁶ Due to the lack of reference English translations, we restrict this setting to X→English and use the reference-free metric MetricX24-QE. As shown in Tables 9, 10, and 11, thinking and non-thinking modes result in very similar performance across languages and metrics. In Chinese, the non-thinking mode slightly outperforms thinking when translating from English, whereas results are nearly identical for Portuguese variants.

We further extend our analysis to document-level translation using TED talks from IWSLT2017 (Cettolo et al., 2012). Sentences are grouped into paragraphs of 12 consecutive sentences, and we evaluate translation between English and French, German, Chinese, and Japanese. Performance is measured using BLEU, XCOMET-XXL, and

⁶<https://github.com/JN-Jeong/Korean-Honorific-Translation>

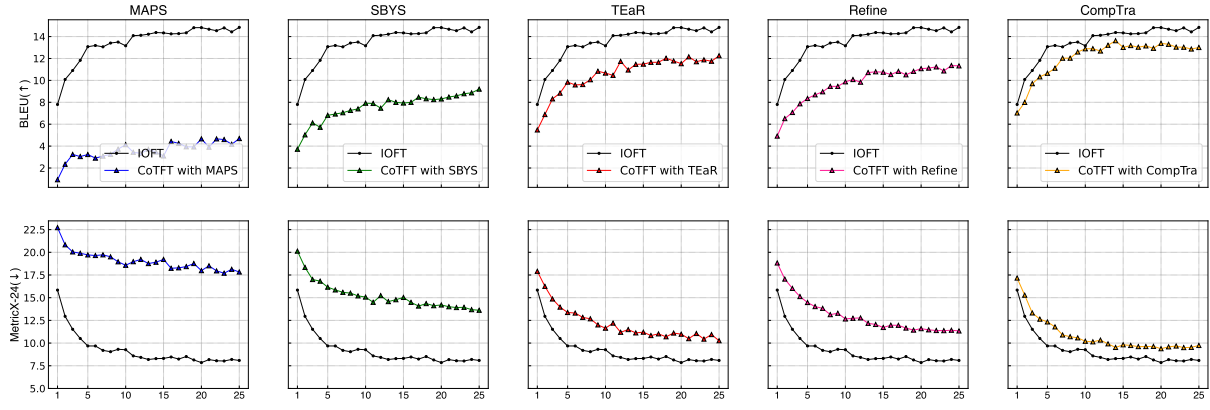


Figure 15: Comparison between IOFT and CoTFT with five different prompting strategies.

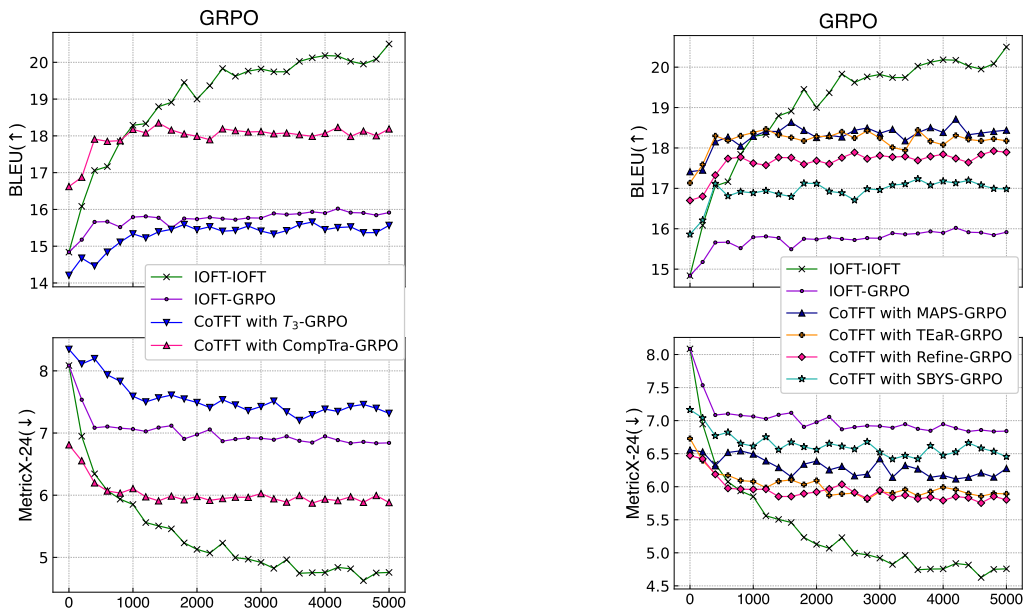


Figure 16: Comparison between IOFT and CoTFT with GRPO.

MetricX24-XXL-QE. As reported in Table 12, the non-thinking mode performs slightly better when translating into English. In the English→X direction, thinking shows small gains for French and German, though the differences remain marginal (0.12 and 0.28 MetricX-QE, respectively). Overall, these results are consistent with our findings in sentence-level and stylized translation: thinking tokens have limited impact on translation quality. That said, we emphasize that these experiments are preliminary. A more thorough investigation, particularly into evaluation methods and broader settings, is needed to fully understand these effects, although our results provide a strong starting point.

C Ablation studies and examples of prompts

C.1 Ablation Studies

In this section, we analyze how our hyperparameter choices influence the observed results. We conduct three main experiments:

- **Increasing the number of IOFT steps.** Rather than training for 5000 steps as described in Section 5.1, we extend IOFT training to 10000 steps to assess whether additional optimization results in further improvements.
- **Increasing gradient accumulation for CoTFT.** This experiment shares the same objective as the first. Beyond simply extending the number of training steps, increasing

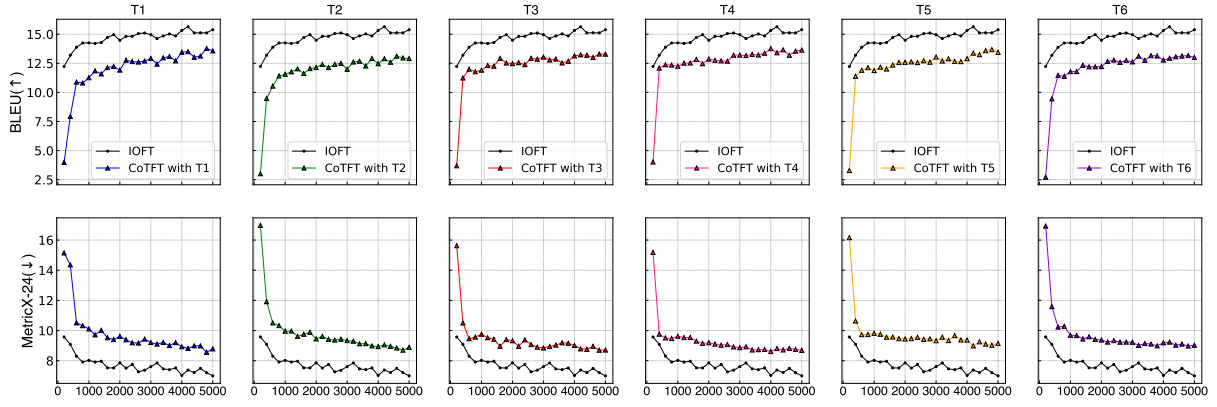


Figure 17: Comparison between IOFT and CoTFT with six different CoT templates (gemma-3-12b-pt).

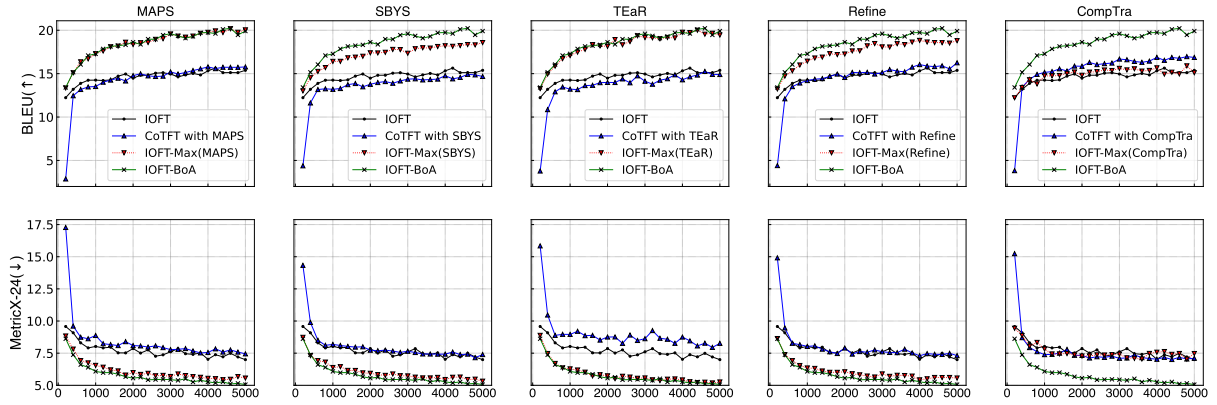


Figure 18: Comparison between IOFT and CoTFT with five different prompting strategies (gemma-3-12b-pt).

English → X	Simplified	Traditional	Yue
<i>w/o thinking</i>			
BLEU	37.32	29.90	34.29
MetricX	1.74	1.95	2.08
<i>w/ thinking</i>			
BLEU	35.39	28.98	32.77
MetricX	1.76	2.01	2.23
X → English	Simplified	Traditional	Yue
<i>w/o thinking</i>			
BLEU	34.80	32.53	36.02
MetricX	1.38	1.51	1.48
<i>w/ thinking</i>			
BLEU	32.56	31.00	34.10
MetricX	1.41	1.58	1.49

Table 9: Stylized translation results for Chinese variants. Best results are highlighted in bold (higher for BLEU and lower for MetricX).

English → X	Portuguese	Brazilian Portuguese
<i>w/o thinking</i>		
BLEU	51.16	54.96
MetricX	1.93	1.82
<i>w/ thinking</i>		
BLEU	51.21	54.49
MetricX	1.88	1.77
X → English	Portuguese	Brazilian Portuguese
<i>w/o thinking</i>		
BLEU	53.47	58.01
MetricX	1.41	1.13
<i>w/ thinking</i>		
BLEU	51.17	56.06
MetricX	1.43	1.14

Table 10: Stylized translation results for Portuguese variants. Best results are highlighted in bold (higher for BLEU and lower for MetricX).

the effective batch size is another way to scale training. Consequently, we study how varying the number of gradient accumulation steps affects CoT distillation.

- **Scaling GRPO.** We adjust the configuration

from Appendix A.5 by running GRPO for 10000 steps and generating 24 samples per prompt.

In addition to these experiments, we evaluate GRPO with an auxiliary reward. Starting from

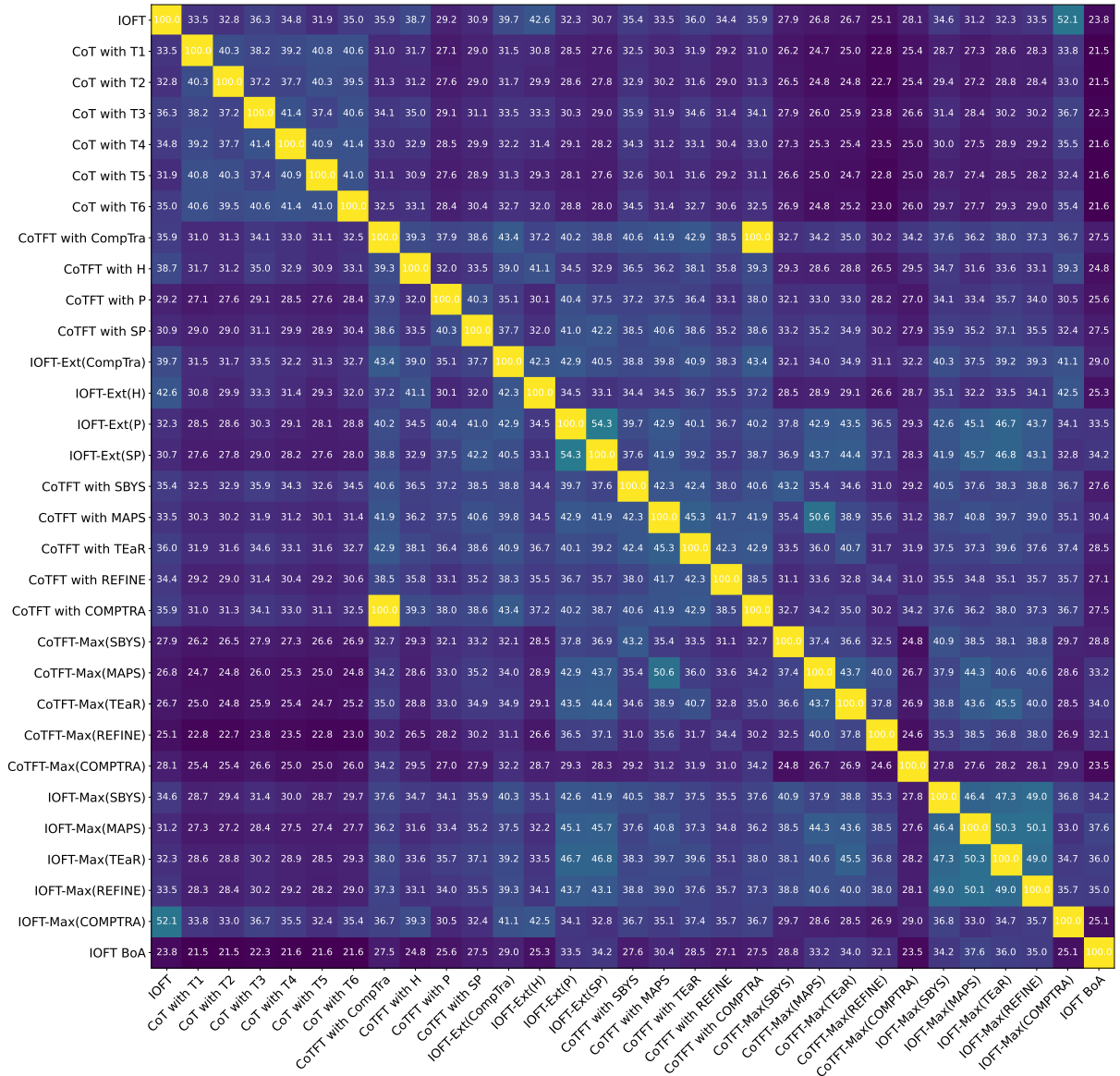


Figure 19: BLEU scores between the FLORES-200 generations of each pair of methods.

Korean → X	Non-honorific	Honorific
<i>w/o thinking</i>		
MetricX24 QE	3.29	3.08
<i>w/ thinking</i>		
MetricX24 QE	3.32	3.00

Table 11: Stylized translation results for Korean honorific variants. Best results are highlighted in bold (lower for MetricX).

a model SFT-trained via CoTFT with CompTra, we introduce a reward based on the average BLASER-2.0 QE score computed between the phrase–translation pairs produced during generation. All other factors are held constant to isolate the effect of this reward on GRPO performance.

Figure 20 summarizes the results of these four experiments. For SFT, increasing training scale provides minimal benefits, indicating that our existing configuration is already adequate. This also suggests that the advantage of IOFT is not due to insufficient CoTFT training. In contrast, scaling RL fine-tuning provides modest improvements. Most of the gains occur early in training (before step 5000), but extending training further still leads to slight additional improvements, particularly on MetricX, while BLEU remains largely unchanged.

We also assess the diversity of generations produced after each fine-tuning strategy. To do so, we compile all strategies into a BLEU-based similarity matrix, where each entry reflects the word-level similarity between the FLORES-200 outputs of a

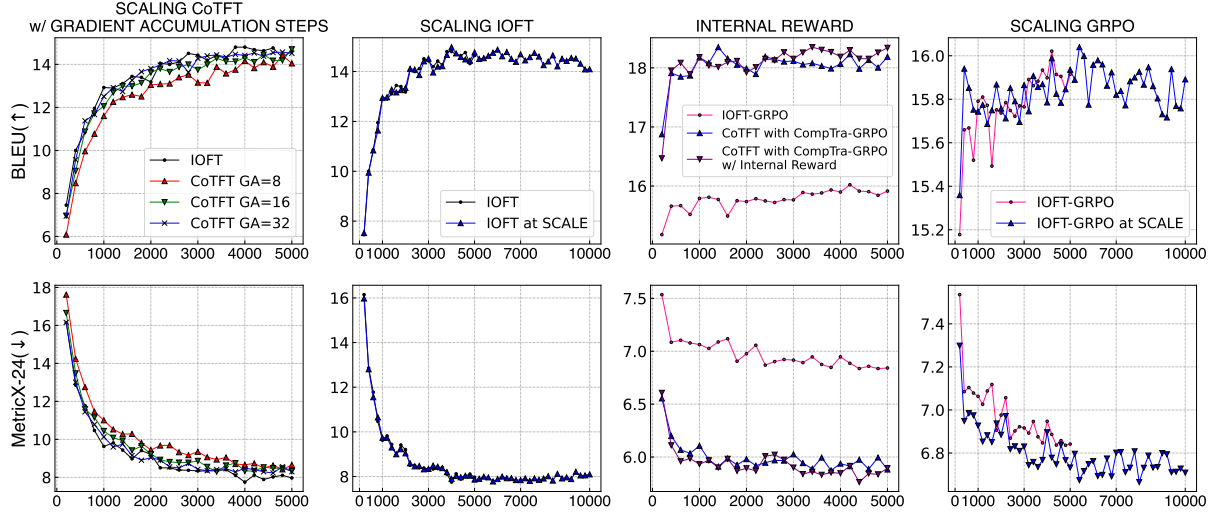


Figure 20: Ablation studies on the hyperparameters. The x-axis represents the number of steps.

English → X	French	German	Chinese	Japanese
<i>w/o thinking</i>				
BLEU	41.54	30.28	20.63	7.83
XCOMET-XXL	28.98	46.30	41.77	30.06
MetricX24-XXL QE	5.75	3.22	3.95	5.89
<i>w/ thinking</i>				
BLEU	43.86	31.03	18.95	7.70
XCOMET-XXL	30.26	45.27	41.62	30.15
MetricX24-XXL QE	5.63	3.40	3.95	5.83
X → English	French	German	Chinese	Japanese
<i>w/o thinking</i>				
BLEU	43.62	35.75	31.38	21.74
XCOMET-XXL	40.85	45.65	37.38	26.74
MetricX24-XXL QE	5.24	3.47	4.40	5.98
<i>w/ thinking</i>				
BLEU	43.00	35.30	28.73	20.75
XCOMET-XXL	40.78	47.39	38.06	27.43
MetricX24-XXL QE	5.25	3.48	4.39	5.88

Table 12: Document-level translation results across languages. Best results are highlighted in bold (higher for BLEU & COMET and lower for MetricX).

pair of methods (applied to gemma-3-4b-pt). As shown in Figure 19, CoT distillation produces more similar generations across templates than any other fine-tuning approach. We observe the same pattern for CoTFT across different prompting strategies. IOFT is most similar to IOFT-MAX(COMPTRA), likely because BLASER-2.0 QE frequently prefers the ground-truth translation over the translation attempts embedded with CompTra’s traces (see Section 4). It is also close to IOFT-EXT(COMPTRA) and IOFT-EXT(H), supporting our intuition that short phrases and hard expressions are not ideal as additional training examples (Section 5.4). Overall, BLEU scores rarely exceed 50, indicating substantial reformulation across strategies. While overlaps do exist, even strategies achieving very similar

MT performance often produce noticeably different generations.

C.2 Examples of training samples

During CoT distillation, we prompt a teacher (e.g. Llama-4-Scout-17B-16E-Instruct) with a source-target pair and ask it to produce a reasoning explaining how to derive the target given the source. Here are some examples with source = “Agdistis was a Phrygian deity known for his extraordinary powers.” and target (in Xhosa) = “Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga.” for each of the CoT templates T1-T6.

T_1

```
<think>
I am analyzing the sentence structure of
the source sentence and identifying
the core elements, which are "
Agdistis" (subject), "was" (verb),
and "a Phrygian deity known for his
extraordinary powers" (object).
I am recognizing that "Agdistis" is a
proper noun and will likely remain
unchanged in the target sentence.
I am identifying the key components of
the object: "Phrygian deity" and "
extraordinary powers".
I am translating "Agdistis" to Xhosa,
which remains "Agdistis" since it's
a proper noun.
I am translating "was" to Xhosa, which
is "yaying" (a form of the verb "to
be" in Xhosa).
I am translating "a Phrygian deity" to
Xhosa, which becomes "umthandazeli
wamaPhrygia" (umthandazeli meaning
deity or god, and wamaPhrygia
meaning of Phrygia).
I am translating "known for his
```

extraordinary powers" to Xhosa, which becomes "owaziwayo ngamandla akhe angaqhelekanga" (owaziwayo meaning known for, ngamandla akhe meaning his powers, and angaqhelekanga meaning extraordinary or unusual). I am combining the translated elements to form a basic translation: "Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga". I am reviewing the translation for basic accuracy and grammatical structure, and it seems to align with the target sentence. I am checking for areas that need further refinement, such as word choice, tense, or word order, but the translation appears to be accurate. I am finalizing the translation, ensuring it retains the original meaning while improving readability, and the translation matches the target sentence. </think>	rough Xhosa sentence: "Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga". I am checking for accuracy: The noun-verb agreement seems correct, and the connectors appear to be in order. I am adjusting the word order to ensure it aligns with Xhosa grammatical structure: The sentence seems to follow a subject-verb-object word order, which is correct for Xhosa. I am reviewing the translation for naturalness, idiomatic use, and overall flow: The sentence seems natural and idiomatic in Xhosa. I am confirming that the translation "Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga" accurately conveys the meaning of the source sentence. </think>
--	---

T_2

<think> I am identifying the basic elements of the source sentence: The subject is "Agdistis", the verb is "was", and the object is "a Phrygian deity known for his extraordinary powers". I am breaking down the object into its components: "a Phrygian deity" is a noun phrase describing the type of deity, and "known for his extraordinary powers" is a dependent clause describing the deity. I am converting these elements into an intermediate language structure: I have "Agdistis" as the subject, "was" as the verb, and "a Phrygian deity" and "known for his extraordinary powers" as descriptors. I am translating "Agdistis" to Xhosa, which remains "Agdistis" since it's a proper noun. I am translating "was" to Xhosa, which can be "yayingumthandazeli" or a similar linking verb, but I will refine this later. I am translating "a Phrygian deity" to Xhosa, which becomes "umthandazeli wamaPhrygia". I am translating "known for his extraordinary powers" to Xhosa, which becomes "owaziwayo ngamandla akhe angaqhelekanga". I am combining these elements into a

T_3

<think> I am analyzing the provided context in the source language, which is the English sentence: "Agdistis was a Phrygian deity known for his extraordinary powers." I identify the key elements in the source sentence: "Agdistis" (proper noun), "Phrygian deity" (noun phrase), and "known for his extraordinary powers" (adjectival phrase describing the deity). I am now translating the source text to the target language, which is Xhosa. I start by translating "Agdistis" to Xhosa, which seems to remain "Agdistis" as it is a proper noun. I then translate "was a Phrygian deity" to Xhosa. "Deity" translates to "umthandazeli" and "Phrygian" to "wamaPhrygia", so this part becomes "yayingumthandazeli wamaPhrygia". Next, I translate "known for his extraordinary powers". "Known for" translates to "owaziwayo ngamandla", and "extraordinary powers" to "angaqhelekanga", so this part becomes "owaziwayo ngamandla akhe angaqhelekanga". I combine these translations to get the target sentence: "Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga".

I perform back translation from the target language to the source language to verify accuracy.

The target sentence "Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga" back-translates to: "Agdistis was a Phrygian deity known for his extraordinary powers".

I compare the back translation with the original source context and see that they match.

I evaluate whether the meaning of the back translation aligns with the original and find that it does.

Since there are no discrepancies identified, I finalize the translation.

I ensure both forward and back translations accurately align across all languages involved.

The translation process confirms that "Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga" is an accurate translation of "Agdistis was a Phrygian deity known for his extraordinary powers".

T_4

<think>

I am analyzing the current sentence, "Agdistis was a Phrygian deity known for his extraordinary powers," along with the fact that there are no previous sentences, to understand that this is a standalone statement about a historical or mythological figure.

I identify that the tone is formal and the subject matter is mythological or historical, indicating a need for a formal and respectful tone in the translation.

I am translating the sentence, starting with the name "Agdistis," which seems to be a proper noun and will likely remain unchanged or transliterated in the target language.

I recognize the phrase "was a Phrygian deity" as a descriptor of Agdistis's origin or classification, which will need to be translated into Xhosa in a way that respects the formal tone and accurately conveys the meaning.

I note the phrase "known for his extraordinary powers," which indicates a need to translate "extraordinary powers" in a way that

is understandable and natural in Xhosa.

I am considering how to translate "Agdistis was a Phrygian deity" into Xhosa, thinking about how to express "deity" and "Phrygian" in the target language, likely using "umthandazeli" for deity and "wamaPhrygia" for Phrygian.

I am translating "known for his extraordinary powers" by thinking about how to say "extraordinary powers" in Xhosa, considering "amandla akhe angaqhelekanga" as a possible translation.

I refine my translation to ensure it fits the context, deciding on "Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga" as a coherent and natural-sounding sentence in Xhosa.

I verify that "Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga" maintains coherence as a standalone sentence and ensures consistency in language and tone, which appears to be formal and informative.

I finalize the translation by cross-checking it with the context, which is essentially nonexistent, and decide that "Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga" feels natural and appropriately aligned for a formal discussion about mythological figures.

T_5

<think>

I am analyzing the source sentence "Agdistis was a Phrygian deity known for his extraordinary powers" to identify the key elements, which include the subject "Agdistis", the verb "was", and the descriptive phrase "a Phrygian deity known for his extraordinary powers".

I am determining that the most suitable translation strategy for this sentence is likely to be a mix of literal and idiomatic translation, as I need to convey the specific cultural and mythological references accurately while also ensuring the sentence flows well in Xhosa.

I am selecting the translation for "Agdistis" as "Agdistis" since proper nouns are often not translated, and it will be retained in its original form in Xhosa.

I am choosing to translate "was" into Xhosa as "yayingum", which is a

linking verb used to describe a state of being, fitting for describing someone's identity or nature. I am translating "a Phrygian deity" into Xhosa as "umthandazeli wamaPhrygia", where "umthandazeli" means deity or god, and "wamaPhrygia" specifies the region or origin, adapting to Xhosa's grammatical structure. I am deciding to translate "known for his extraordinary powers" as "owaziwayo ngamandla akhe angaqhelekanga", where "owaziwayo" means known, "ngamandla" means powers, "akhe" signifies possession (his), and "angaqhelekanga" means extraordinary or unusual, ensuring the descriptive phrase conveys the intended meaning. I am reviewing my translation decisions and making adjustments for fluency and accuracy, ensuring that the Xhosa sentence "Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga" accurately reflects the source sentence. I am explaining that I chose specific words or phrases to ensure that the translation is not only accurate but also natural-sounding in Xhosa, balancing between literal translation and contextual appropriateness. I am providing a final explanation for my translation choices, noting that while "Agdistis" remains unchanged, other parts of the sentence were adapted to fit Xhosa grammar and vocabulary, ensuring the target sentence effectively communicates the information from the source sentence. I am concluding that my translation decisions aimed to maintain the original meaning while making necessary adjustments for linguistic and cultural differences between English and Xhosa. </think>	predicate nominative. I am determining the most appropriate syntactic structure in the target language, and I think I can maintain a similar structure in Xhosa, but I need to consider the language's own grammatical rules and word order. I am adjusting the word order and grammatical structure in the target language to match the sentence's meaning, while maintaining clarity, and I note that Xhosa typically uses an SVO (Subject-Verb-Object) word order. I am translating the sentence, ensuring that subject-verb-object relationships and other syntactic elements align with target language norms, and I start with "Agdistis" as the subject, which becomes "Agdistis" in Xhosa. I am translating the linking verb "was" and consider using the Xhosa equivalent "yayingum" (a form of "to be"). I am translating the predicate nominative "a Phrygian deity" and consider using "umthandazeli wamaPhrygia" (a Xhosa phrase for "a Phrygian deity"). I am translating the phrase "known for his extraordinary powers" and consider using "owaziwayo ngamandla akhe angaqhelekanga" (a Xhosa phrase that maintains the same meaning). I am checking the sentence's grammar and overall flow in the target language, and I have "Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga". I am refining the structure by adjusting word choice or reordering components to ensure the sentence feels natural and fluid in Xhosa, and I think "Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga" meets these criteria.
T_6 <think> I am analyzing the sentence's syntactic structure in the source language, and I see that "Agdistis was a Phrygian deity known for his extraordinary powers" is in the passive voice, with "Agdistis" as the subject, "was" as the linking verb, and "a Phrygian deity known for his extraordinary powers" as the	I am confirming that the translation accurately conveys the original meaning and is clear and fluid in the target language, and I am satisfied with the result. </think> For MT prompting strategies, we first prompt the teacher with the source sentence to generate all intermediate steps required by each strategy. We then

concatenate these steps into a single training sample while keeping the original ground-truth target unchanged. Below, we provide example training samples for the five strategies used in this work.

SBYS

<think>
During the research phase, I have identified the following potential translation challenge:

* Idiomatic Expressions:

There are no apparent idiomatic expressions in the given context text: "Agdistis was a Phrygian deity known for his extraordinary powers." However, the phrase "extraordinary powers" may pose a challenge as it is an idiomatic expression that may not have a direct equivalent in Xhosa.

In English, "extraordinary powers" is an idiomatic expression used to describe exceptional or supernatural abilities. To translate this phrase into Xhosa, I may need to use a more literal translation or find an equivalent Xhosa expression that conveys the same meaning.

Additionally, I have noted that:

* "Phrygian" refers to an ancient region in Asia Minor, and the term may not be well-known in Xhosa culture. This may require some explanatory notes or a more detailed translation to convey the cultural and historical context.

No other idiomatic expressions or potential translation challenges were identified in the given context text.

Please let me know if I should proceed with the next stage of the translation process.

UAgdistis waye uThixo wasePhrygia owaziwa ngamandla akhe angaqhelekanga.

Here's a brief explanation of my translation choices:

* "Agdistis" is left untranslated as it is a proper noun and a name of a deity.

* "was a" is translated to "waye u" which is a Xhosa copular phrase used to describe a person's identity or status.

* "Phrygian" is translated to "wasePhrygia" which indicates the region or origin of the deity.

* "deity" is not directly translated,

but implied in the context of "UThixo" which means "God" or "deity".

* "known for" is translated to "owaziwa ngamandla" which means "known for his powers".

* "extraordinary powers" is translated to "angaqhelekanga" which is an adverb that means "unusual", "exceptional" or "remarkable", and "amandla" which means "powers".

Please let me know if I should proceed with the next stage of the translation process.

Now let's move to the next stage: Post-editing with local refinement. In this stage, the primary aim is to refine the draft translation by making micro-level improvements that improve the draft's fluency.

Here is a refined version of the translation

UAgdistis waye uThixo wasePhrygia owaziwangamandla akhe ayengavamile.

Now, we will proofread the refined text for grammar spelling, punctuation, terminology and overall fluency.

Here is the translation after proofreading

After proofreading the refined translation, I suggest the following final, polished translation:

We will further improve it to obtain the final, polished translation.

</think>

Final Translation

Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga.

MAPS

<think>
Here is a draft translation

1. UAgdistis waye uhlambe wasePhrygia owaziwa ngamandla akhe angaqhelekanga.

Let's write an English sentence related to but different from the input English sentence and translate it into Xhosa

The cult of Agdistis, a Phrygian deity, was centered there, and the site was probably a major religious center. Isikhululo sika Agdistis, isithixo sasePhrygian, sasigxile apho, yaye indawo yayo yayisikhungo esikhulu senkonzo yezokholo.

Given this knowledge, we can draft

another translation 2. Agdistis wayesithixo sasePhrygian esaziwa ngamandla akhe angaqhelekanga. Let's extract the keywords in the provided English sentence, and then translate these keywords into Xhosa * Agdistis=Agdistis * Phrygian=Frighiya * deity=ility * extraordinary=angakhethekileyo * powers=amandla Given this knowledge, we can draft another translation 3. UAgdistis wayesithixo saseFrighiya esaziwa ngamandla akhe angakhethekileyo. Let's use a few words to describe the topics of the provided English sentence Mythology, ancient deity Given this knowledge, we can draft another translation 4. Agdistis wayeyona dea yasePhrygian eyaziwa ngamandla ayo angaqhelekanga . We will choose the best of these translations and further improve it to obtain the final, polished translation. </think> Final Translation Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga.	
--	--

TEaR

<think> Here is a draft translation 1. Agdistis wayesithixo sasePhrygia esaziwayo ngamandla akhe angaqhelekanga. Let's identify errors and assess the quality of the draft translation. The categories of errors are accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation, register , spelling), locale convention (currency, date, name, telephone, or time format) style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, other, or no-error. Each error is classified as one of three	
--	--

categories: critical, major, and minor. Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what the text is trying to say is still understandable. Minor errors are technical errors but do not disrupt the flow or hinder comprehension. Here are the MQM annotations of the draft: critical: no-error major: no-error minor: no-error Upon reviewing the translation and error information, we can refine the draft and obtain a better translation 2. Agdistis wayesithixo sasePhrygia esaziwayo ngamandla akhe angavamile. We will further improve it to obtain the final, polished translation. </think> Final Translation Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga.	
---	--

Self-Refine

<think> Here is a draft translation 1. UAgdistis waye uhlambe wasePhrygia owaziwa ngamandla akhe angaqhelekanga. Let's improve it and write a better translation 2. UAgdistis waye uhlambe wasePhrygia owaziwa ngamandla akhe amakhulu. Let's further improve it and write a better translation 3. UAgdistis waye uhlambe wasePhrygia owaziwa ngamandla akhe angaqhelekanga. Let's improve it one last time and write a better translation 4. UAgdistis waye uhlambe wasePhrygia owaziwa ngamandla akhe amakhulu. We will choose the best of these translations and further improve it to obtain the final, polished translation. </think> Final Translation Agdistis yayingumthandazeli wamaPhrygia owaziwayo ngamandla akhe angaqhelekanga.	
---	--

CompTra

<Demonstrations>

1. English sentence

Agdistis was a Phrygian deity.

Xhosa translation

UAgdistis wayesithixo sasePhrygia.

2. English sentence

Agdistis was known for his extraordinary powers.

Xhosa translation

Agdistis waziwa ngamalawulo akhe
angaqhelekanga.

</Demonstrations>

Final Translation

Agdistis yayingumthandazeli wamaPhrygia
owaziwayo ngamandla akhe
angaqhelekanga.