Compositional Translation: A Novel LLM-based Approach for Low-resource Machine Translation

Armel Zebaze **Benoît Sagot** **Rachel Bawden**
Inria, Paris, France
firstname.lastname@inria.fr

Abstract

The ability of generative large language models (LLMs) to perform in-context learning has given rise to a large body of research into how best to prompt models for various natural language processing tasks. Machine Translation (MT) has been shown to benefit from in-context examples, in particular when they are semantically similar to the sentence to translate. In this paper, we propose a new LLM-based translation paradigm, *compositional translation*, to replace naive few-shot MT with similarity-based demonstrations. An LLM is used to decompose a sentence into simpler phrases, and then to translate each phrase with the help of retrieved demonstrations. Finally, the LLM is prompted to translate the initial sentence with the help of the self-generated phrase-translation pairs. Our intuition is that this approach should improve translation because these shorter phrases should be intrinsically easier to translate and easier to match with relevant examples. This is especially beneficial in low-resource scenarios, and more generally whenever the selection pool is small or out of domain. We show that compositional translation boosts LLM translation performance on a wide range of popular MT benchmarks, including FLORES-200, NTREX 128 and TICO-19. Code and outputs are available at https://github.com/ArmelRandy/compositional-translation.

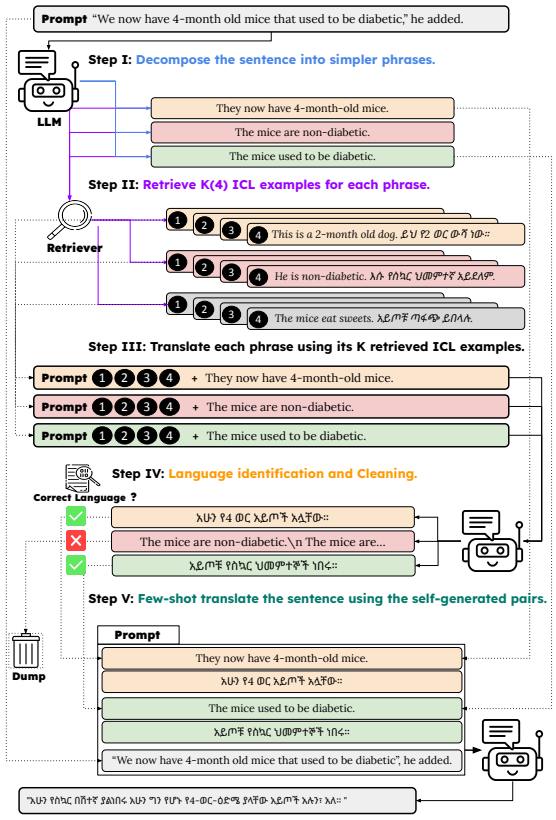

Figure 1: An overview of compositional translation (CompTra). Given a sentence, we prompt the LLM to decompose it into several phrases that use its words. For each phrase, we retrieve relevant in-context demonstrations (here four) through similarity search and use them to translate it in a few-shot setup. The phrase-translation couples obtained are then cleaned and provided to the LLM to help it translate the main sentence.

1 Introduction

Large Language Models (LLMs) have demonstrated strong performance across a wide variety of tasks (Chowdhery et al., 2023; Dubey et al., 2024). They use In-Context Learning (ICL; Brown et al., 2020) to solve problems at inference with the help of a few examples within their context. Multiple strategies have been introduced to expand the range of complexity of problems that can be addressed using ICL, with a particular emphasis on reasoning tasks (Wei et al., 2022; Kojima et al., 2022; Yao et al., 2023). In MT, most works have focused on choosing the best in-context examples. The prevailing intuition is to choose them based on their word-level similarity to the source sentence (Agrawal et al., 2023; Zhang et al., 2023a; Moslem et al., 2023; Vilar et al., 2023; Zhu et al., 2024; Bouthors et al., 2024). This intuition implicitly relies on the

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22328–22357
November 4-9, 2025

property of compositionality of MT (Turcato and Popowich, 2001); the translation of a sequence of words can be modeled as a function of the translation of its subparts. Choosing in-context examples based on similarity search amounts to finding subparts of a sentence to translate in another sentence that has a known translation.

While most studies only marginally explore LLMs for MT into low-resource languages (LRLs), we place them at the center of this work for two reasons: (i) although current LLMs have narrowed the gap with supervised MT models for high-resource language (HRL) directions, they still struggle when translating into LRLs (Hendy et al., 2023; Enis and Hopkins, 2024), and (ii) a few works have shown that example selection based on similarity can improve translation into LRLs (Moslem et al., 2023; Tanzer et al., 2024; Zebaze et al., 2024a), suggesting a promising avenue for further research. We propose compositional translation (CompTra), an LLM-based MT paradigm where translation is explicitly done step by step via the decomposition of the sentence into shorter and simpler entities. Puduppully et al. (2023) proposed DecoMT as a decomposition-based approach to MT, in which a sentence is divided into subparts at the token-level and the translation is obtained by progressively solving fill-in-the-middle tasks. However their decomposition is uninformed, resulting in incongruous subparts hard to translate, and the sequential nature of their approach makes it slow and suboptimal for decoder-based models. Similarly, Ghazvininejad et al. (2023) and Lu et al. (2024) proposed to extract keywords from a sentence and translate them respectively with a dictionary and a multilingual dictionary. While the pieces obtained after decomposition are sound, their sizes do not provide enough context for accurate translation using an LLM. Moreover, their reliance on a dictionary considerably limits the impact of the intrinsic capabilities of the LLM on the MT task.

In CompTra, given a sentence to translate, we prompt an LLM to generate simpler, concise and independent phrases that capture some of its aspects and use its words in the same context. We then proceed to self-generate their translations in a few-shot fashion, with demonstrations retrieved via similarity search in a selection pool (tightening the constraint on the access to outside knowledge). Finally, the LLM derives the main translation by drawing insights from the self-generated translation pairs. The underlying idea is that (i) LLMs are more effective at handling short phrases, and (ii) it is easier to retrieve relevant in-context demonstrations for translating such segments. These phrase translations will be of higher quality, and the similarity between the phrases and the source sentence will enable LLMs to produce better translations. When the selection pool is small and lacks diversity, these self-generated phrase-translation pairs ensure high-quality and similar in-context demonstrations for the main translation. This contrasts with example selection via similarity search, which may not always provide the same level of similarity. We evaluate CompTra on MT from English to LRLs: 10 languages from FLORES-200 (Goyal et al., 2022; Costa-jussà et al., 2022) and 10 languages from NTREX-128 (Federmann et al., 2022) and TICO-19 (Anastasopoulos et al., 2020). Our experiments with `Command-R+`, `LLaMA-3.1-70B-It` and `Gemma-2-27B-It` show that CompTra outperforms example selection via similarity search and many existing strategies.

2 Related Work

Using LLMs for Machine Translation Similar to natural language understanding, reasoning and code generation, LLMs have been successfully applied to MT. Brown et al. (2020) showed that GPT-3 few-shot performance was on par with the state of the art for some translation directions at the time of its release. Bawden and Yvon (2023), Vilar et al. (2023) and Zhang et al. (2023a) explored aspects of few-shot MT including template selection and in-context example selection with BLOOM (BigScience Workshop et al., 2023), PaLM (Chowdhery et al., 2023) and GLM 130B (Zeng et al., 2023) respectively. Agrawal et al. (2023), Mu et al. (2023) and Bouthors et al. (2024) documented performance gains when choosing in-context demonstrations semantically related to the sentence to translate. Hendy et al. (2023) demonstrated that GPT models perform well as zero- and few-shot translators but face challenges with LRL pairs. Building on this, Zhu et al. (2024) conducted an extensive analysis across 102 languages, confirming these findings. They also noted that similarity search in a high-quality pool offers no significant benefit. In contrast, Zebaze et al. (2024a) showed that, while similarity-based selection does not help HRLs, it significantly improves performance for LRLs.

Prompting and Compositionality ICL was shown to work for diverse tasks (Brown et al.,

2020), leading to the development of new problem solving strategies at inference. Wei et al. (2022) introduced chain-of-thought (CoT) prompting, helping LLMs to mimic a step-by-step thought process by providing reasoning steps in the demonstrations. Following this, multiple works emerged on the necessity of in-context examples for CoT prompting (Kojima et al., 2022) and on their design (Zhang et al., 2023b; Fu et al., 2023; Yasunaga et al., 2024). More advanced techniques include self-consistency (Wang et al., 2023) and hierarchical approaches such as Tree of Thoughts (ToT) (Yao et al., 2023) and Graph of Thoughts (GoT) (Besta et al., 2024). Another line of research involves teaching LLMs to tackle complex problems by breaking them down into a series of subproblems and recursively solving them to derive the final answer (Dua et al., 2022; Zhou et al., 2023; Khot et al., 2023; Zebaze et al., 2024b). All these efforts consistently enhanced the reasoning abilities of LLMs but had a limited effect on MT.

Prompting LLMs for Machine Translation Beyond example selection for few-shot MT, several works have proposed prompting strategies for MT. Puduppully et al. (2023) proposed DecoMT, which decomposes a sentence to translate into chunks of tokens, independently translates them, and derives the final translation by contextually translating each chunk one after another. The contextual translation of a chunk is obtained by using the contextual translation of the previous chunk as the left context and the independent translation of the next chunk as the right context. This inherently limits DecoMT's applicability to models trained like T5 (Raffel et al., 2020) or trained with the Fill-In-the-Middle (FIM) objective (Bavarian et al., 2022). While our work shares the idea of decomposition, we seek to derive simple, well-formed and coherent phrases that can be accurately translated independently from each other and directly used in few-shot for the main translation, making our approach non-sequential. We propose a decomposition into subparts depending on the structure of the sentence (thus hyperparameter-free) where words in common with the main sentence are used in the same context with the end-goal of leveraging the property of compositionality of MT. Ghazvininejad et al. (2023) introduced Dictionary-based Prompting for MT (DiPMT), which uses a dictionary to provide the target translations of certain words within a source sentence. These translations are incorporated into the input to help the LLM generate better translations. Building on this idea, Lu et al. (2024) proposed Chain-of-Dictionary (CoD), which extends DiPMT by translating chunks of words into multiple auxiliary languages and the target language using a multilingual dictionary, such as NLLB (Costa-jussà et al., 2022). These translations are provided as additional context to further improve translation quality. In contrast, our target is to improve LLM-based translation quality by only relying on the LLM itself.

Another line of research involves progressively guiding LLMs to produce good translations through a self-refinement process with or without external feedback (Chen et al., 2024; Feng et al., 2024; Xu et al., 2024d; Ki and Carpuat, 2024) inspired by the success of this paradigm for reasoning tasks (Madaan et al., 2023; Shinn et al., 2024). This refining step is included in many strategies for MT. Briakou et al. (2024) proposed "Translating Step-by-Step" (SBYS): a multi-turn interaction with an LLM that breaks down the translation process into four distinct stages: identification of challenging components, drafting, refinement and proofreading. Feng et al. (2024) designed another multi-step approach, "Translate, Estimate, and Refine" (TEaR), where a model generates a draft translation, self-derives the MQM annotations of the draft with the help of few-shot examples and subsequently refines the translation based on these annotations. On a different note, He et al. (2024) proposed "Multi-Aspect Prompting and Selection" (MAPS), an ensembling technique that involves prompting a LLM to analyze a sentence for translation by building knowledge across three key aspects: keywords, topics, and relevant demonstrations. Each aspect guides the LLM in generating a candidate translation. The final translation is then selected from these three candidates, along with the zero-shot output, based on the highest COMET QE (Rei et al., 2020) score relative to the source sentence.

LLMs and Low-Resource Languages LLMs are trained on increasingly larger datasets in accordance with scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022; Dubey et al., 2024), which has led to them becoming more multilingual (BigScience Workshop et al., 2023; Cahyawijaya et al., 2024; Enis and Hopkins, 2024), whether intentionally or not. Fan et al. (2021) and Costa-jussà et al. (2022) developed supervised multilingual MT models that significantly improved the state-

of-the-art for various LRLs. Building on these advancements, subsequent progress included the release of massively multilingual datasets (Schwenk et al., 2021; Abadji et al., 2022; ImaniGooghari et al., 2023; Singh et al., 2024; Futeral et al., 2024), as well as efforts in multilingual and multi-task fine-tuning (Muennighoff et al., 2023; Üstün et al., 2024; Lai et al., 2024) and continual pre-training (Xu et al., 2024a,c,b; Dou et al., 2024).

3 Methodology

We introduce Compositional Translation (CompTra), an LLM-based translation paradigm that automatically allows LLMs to translate any sentence by reasoning on self-generated translation pairs tailored to its content. CompTra frames any translation problem as an explicit step-by-step procedure. It consists of three main stages.

1. **Decomposition.** Given that the use of related in-context examples helps few-shot MT into LRLs, we want to derive pairs as closely related to the source sentence as possible. The aim of decomposition is to create simpler phrases that share words with the source sentence. We achieve this using a `divide prompt`, which contains examples that demonstrate the decomposition, followed by the source sentence. The examples are from the MinWikiSplit corpus (Niklaus et al., 2019), a set of sentences broken down into minimal propositions. It is worth noting that the number of phrases obtained is not a hyperparameter; each sentence is decomposed into the number of phrases that fits its structure.

2. **Translation.** The LLM independently translates each of the phrases obtained. The `translate prompt` uses some artifacts, typically few-shot examples chosen via similarity search with a retriever. In practice, the phrases' translations are often written in an incorrect target language. We filter out phrase translations in the incorrect language with the help of a language identifier.

3. **Recombination.** The LLM is fed with the phrases obtained after decomposition and their self-generated translations combined into a `merge prompt`. In our experiments, this prompt has exactly the same structure as the `translate prompt` in order to decouple the gains seen from changes to the prompt.

The only hyperparameter is the number of demonstrations per phrase k, which we set to 5 for all phrases. The hypothesis is that LLMs translate short sentences more effectively than longer ones, especially in languages they marginally encountered during their training. CompTra's objective is to propagate this advantage of short entities to bigger ones via a three-step hierarchical approach.

4 Experiments

Datasets from English to LRLs.

- **FLORES-200** (Goyal et al., 2022; Costa-jussà et al., 2022). This dataset consists of translations from web articles into 204 languages. We use its dev set (997 examples) as the selection pool and its devtest set (1012 examples) for evaluation.

- **NTREX 128** (Federmann et al., 2022; Barrault et al., 2019) is an MT benchmark derived from WMT19 news data translated by professional human translators. It contains 1997 parallel sentences 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 in 35 languages. Its validation and test sets consist of 971 (used as a selection pool) and 2100 samples respectively.

Models. We use `LLaMA-3.1-It` (8B, 70B; Dubey et al., 2024), `Gemma-2-It` (9B, 27B; Gemma Team et al., 2024), `Command-R` and `-R+`[1] (Cohere, 2024).

Evaluation. We mainly evaluate using MetricX-23 (Juraska et al., 2023) and XCOMET (Guerreiro et al., 2024) in their reference-based versions: `XCOMET-XXL` (which supports the same 100 languages as XLM RoBERTa (Conneau et al., 2020)) and `MetricX-23-XXL` (which supports the same 101 languages as mT5 (Xue et al., 2021)). MetricX scores range from 0 to 25, with higher scores indicating more translation errors. XCOMET scores range from 0 and 1, which we rescale to 0 to 100 (higher scores are better). We also consider n-gram matching metrics via sacreBLEU (Post, 2018),

[1] `command-r-08-2024`, `command-r-plus-08-2024`

namely chrF++[2] (Popović, 2017) and BLEU[3] (Papineni et al., 2002) for transparency reasons.

Baselines. We compare to zero-shot and 5-shot MT with example retrieval via similarity with BM25 (Robertson et al., 1995) and SONAR (Duquenne et al., 2023). The former retrieves the five most similar examples based on its BM25 score relative to the query, whereas the latter uses the cosine similarity between the query and the candidates in the embedding space.

Experimental Details. In all experiments, CompTra uses BM25 (Robertson et al., 1995) as its retriever and queries the five most similar in-context examples for each phrase unless specified otherwise. Language identification is done with FastText (Bojanowski et al., 2017; Costa-jussà et al., 2022) and only when the language is supported. We use `vLLM` (Kwon et al., 2023) for inference with greedy decoding and BM25s (Lù, 2024). We generate at most 500 new tokens during the translation phase and 2000 during combination. We remove repeating bigrams at the end of the translations.[4] We use paired bootstrap resampling (Koehn, 2004) with 300 samples of 500 sentences and a significance threshold of $p < 0.05$.

4.1 Results

4.1.1 Main results: FLORES-200

We evaluate CompTra on 10 English-to-X translation directions from FLORES-200 (Figure 2). The dark unpatterned bars (CompTra) are usually the shortest, indicating that CompTra generally outperforms similarity-based few-shot MT across all directions and for all LLMs evaluated. On average, CompTra outperforms few-shot BM25 by 0.4 MetricX points with `LLaMA-3.1-70B-It` and `Gemma-2-27B-It`, and by 1.5 MetricX points with `Command-R+`. For XCOMET, the gains are 1.0 with `Gemma-2-27B-It` and 1.8 with `Command-R+`. We report the exact numerical scores in Appendix B.1.[5]

4.1.2 Results on NTREX 128 and TICO-19

We report the results obtained by CompTra on NTREX 128 and TICO-19 in Figure 3. Similar to our observations on FLORES-200, CompTra mostly outperforms similarity-based few-shot MT across all directions and for all the LLMs. We report the exact numerical scores in Appendix B.4.[6]

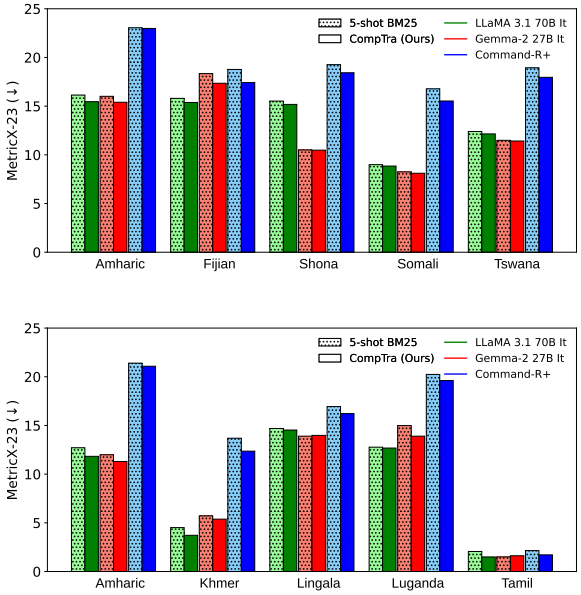

Figure 3: MetricX results for 5 English→X directions from TICO-19 (Anastasopoulos et al., 2020) (above) and NTREX 128 (Federmann et al., 2022) (below).

4.1.3 Comparison to existing approaches

We conduct a set of additional studies and compare[7] CompTra against existing methods including zero- and few-shot MT and CoT (Kojima et al., 2022; Peng et al., 2023), MAPS (He et al., 2024), TEaR (Feng et al., 2024), SBYS (Briakou et al., 2024) and standalone Self-Refine (Chen et al., 2024). We use `LLaMA-3.1-70B-It` and report the results in Figure 4. For each language, the lowest marker indicates the best-performing strategy, while the highest denotes the worst. CompTra significantly outperforms CoT, MAPS, TEaR and SBYS with gains ranging from 0.4 to 3.3 MetricX points. 5-shot BM25 emerges as a very strong baseline, which can be further improved via self-refine, although it does not reach CompTra's performance. CoT degrades the zero-shot and few-shot performance of the model, as opposed to what happens on reasoning tasks. We report the exact numerical scores in Appendix B.2.[8]

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

[3] nrefs:1|case:mixed|eff:no|tok:flores200|smooth:exp|version:2.4.2

[4] See Appendix C.1 for details.

[5] See Appendix B.3 for BLEU and chrF++ scores.

[6] See Appendix B.5 for BLEU and chrF++ scores.

[7] See Appendix A for a runtime comparison.

[8] See Appendix B.3 for BLEU and chrF++ scores.

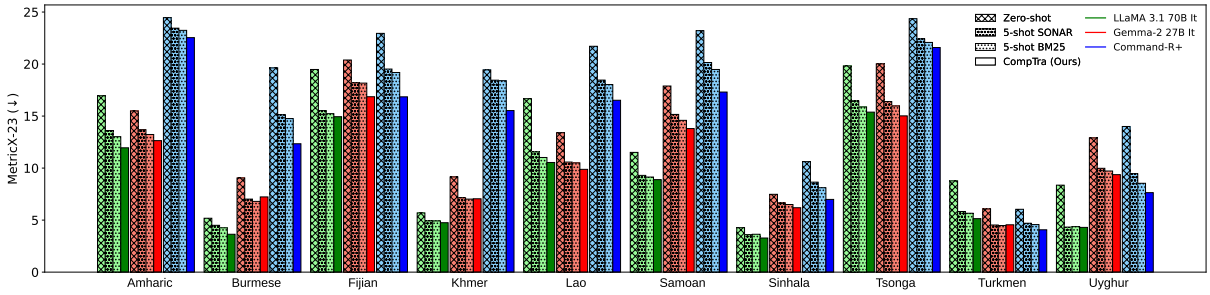

Figure 2: MetricX scores for 10 English→X directions from FLORES-200 (lower is better).

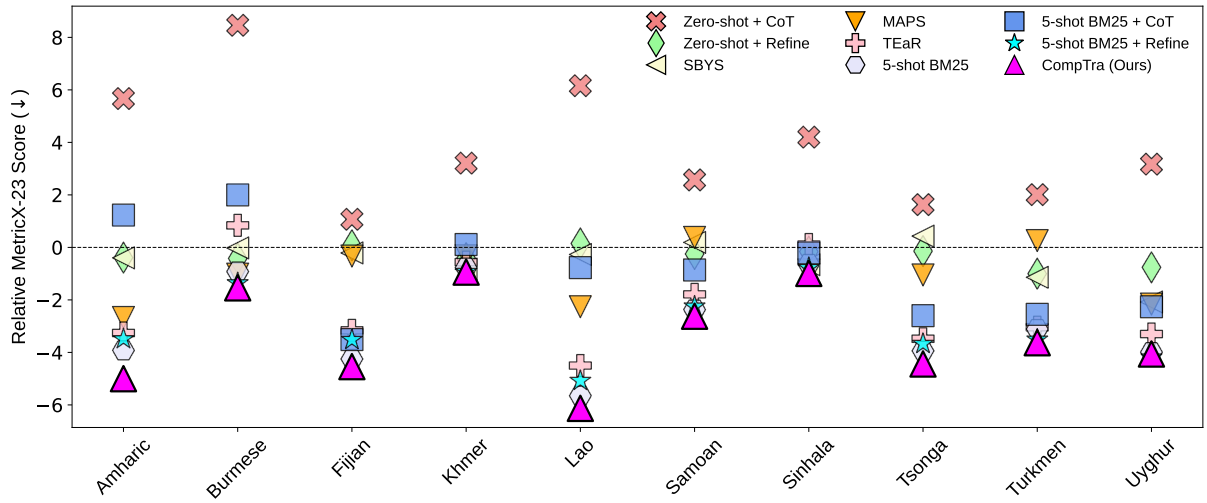

Figure 4: Relative MetricX scores with respect to zero-shot for 10 English→X directions from FLORES-200 (lower is better). We compare CompTra to CoT (Kojima et al., 2022), MAPS (He et al., 2024), SBYS (Briakou et al., 2024) and TEaR (Feng et al., 2024).

5 Analysis

Does CompTra work with weaker LLMs? In all our experiments in Section 4.1, we mainly used `LLaMA-3.1-70B-It`, but can CompTra work with weaker models? We compared CompTra with few-shot BM25 on FLORES-200 when both prompting approaches use the same weaker base LMs `LLaMA-3.1-8B-It`, `Gemma-2-9B-It` and `Command-R`. In Table 3 (See Appendix B.7 for BLEU and chrF++ scores), we observe that CompTra does work with small LMs; the average absolute performance gap across the ten FLORES languages is 1.04 MetricX points with `Gemma-2-9B-It` vs. 0.44 with `Gemma-2-27B-It`; 1.04 with `Command-R` vs. 1.5 with `Command-R+` and 0.4 for both `LLaMAs`. CompTra's simplicity (it does not require LMs to follow complex instructions), makes it applicable at scale.

Out-of-domain evaluation. In previous experiments, the selection pool shared the same domain as the evaluation set. Here, we investigate whether the gains observed disappear when it is no longer the case. To test this, we consider the setup where the evaluation set is the TICO-19 test set (COVID-19; health domain) and the selection pool is the FLORES-200 dev set (news domain) as opposed to the usual TICO-19 validation set (health domain). As expected and reported in Table 5, in-domain scores are better than their out-of-domain counterparts. However, in both scenarios, applying CompTra gives gains over standalone retrieval-based few-shot MT. This suggests that CompTra can be successfully applied in setups where there is a mismatch between the domain of the selection pool and the evaluation set.

What happens when we modify the translation step? In CompTra, the phrases obtained after decomposition are translated by the LLM in a few-shot manner with the help of in-context demonstrations retrieved with similarity search. In this section, we study two setups. First, we analyze the impact of the number of in-context demonstrations

	Amharic	Burmese	Fijian	Khmer	Lao	Samoan	Sinhala	Tsonga	Turkmen	Uyghur
NLLB-200-distilled-600M	**5.46**	6.32	**9.69**	9.92	**5.75**	6.25	5.16	**8.78**	9.04	7.17
LLaMA-3.1-70B-Instruct										
NLLB + CompTra	5.58	5.06	10.12	6.72	**5.75**	**5.54**	**3.19**	**8.74**	6.21	5.66
5-shot BM25	13.02	4.26	15.23	4.92	11.02	9.14	3.64	15.88	5.66	4.37
CompTra with BM25	11.95	**3.64**	14.94	**4.75**	10.54	8.89	3.28	15.38	**5.14**	**4.30**
CoD (with NLLB-200-600M)	15.41	16.35	22.41	14.22	13.26	15.88	16.16	22.36	19.33	14.83

Table 1: Impact of switching few-shot MT with NLLB in CompTra (MetricX scores).

	Amharic	Burmese	Fijian	Khmer	Lao	Samoan	Sinhala	Tsonga	Turkmen	Uyghur
LLaMA-3.1-70B-Instruct										
5-shot SONAR	13.61	4.50	**15.52**	4.92	**11.55**	**9.30**	3.60	**16.44**	5.83	**4.33**
CompTra with SONAR	**12.93**	**4.00**	15.86	4.98	11.95	9.86	**3.27**	16.86	**5.67**	4.79
5-shot LCS	14.74	6.42	16.89	5.40	**13.07**	**10.14**	3.97	**17.58**	6.33	**4.79**
CompTra with LCS	**14.65**	**4.22**	**16.74**	**5.36**	13.44	10.57	**3.59**	17.82	**6.07**	5.07
5-shot BM25	13.02	4.26	15.23	4.92	11.02	9.14	3.64	15.88	5.66	4.37
CompTra with BM25	**11.95**	**3.64**	**14.94**	**4.75**	**10.54**	**8.89**	**3.28**	**15.38**	**5.14**	**4.30**

Table 2: Ablation study on the impact of the retriever on CompTra (MetricX scores).

	Amharic	Burmese	Fijian	Khmer	Lao
LLaMA-3.1-8B-It					
5-shot BM25	23.40	**14.27**	21.74	12.63	22.81
CompTra (Ours)	**23.06**	**14.29**	**20.93**	**12.02**	**22.41**
Gemma-2-9B-It					
5-shot BM25	15.99	13.05	20.66	11.92	15.21
CompTra (Ours)	**15.66**	**12.31**	**19.63**	**11.23**	**13.67**
Command-R					
5-shot BM25	**24.38**	20.94	21.24	21.64	22.68
CompTra (Ours)	**24.39**	**19.33**	**20.59**	**20.48**	**21.88**
	Samoan	Sinhala	Tsonga	Turkmen	Uyghur
LLaMA-3.1-8B-It					
5-shot BM25	19.80	13.79	23.02	14.72	**14.01**
CompTra (Ours)	**18.25**	**13.23**	**22.75**	**14.39**	15.00
Gemma-2-9B-It					
5-shot BM25	17.61	9.13	20.99	8.36	21.07
CompTra (Ours)	**15.93**	**8.82**	**19.82**	**7.69**	**19.19**
Command-R					
5-shot BM25	21.67	15.50	22.46	7.00	16.36
CompTra (Ours)	**20.82**	**12.91**	**22.16**	**5.95**	**14.99**

Table 3: Full MetricX results for ten English→X directions with smaller LMs.

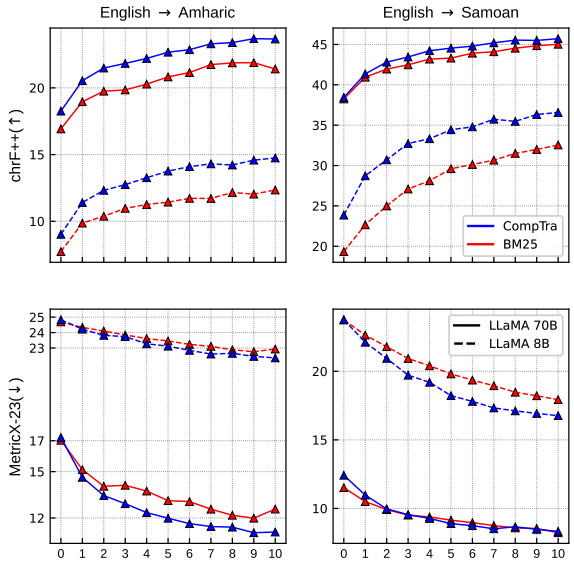

Figure 5: Impact of the number of in-context examples per phrase.

per phrase. As shown in Figure 5, CompTra outperforms few-shot with BM25 as we vary the number of in-context demonstrations and also at scale. The performance gap is as high as 6 chrF++ points and 2.5 chrF++ points in Samoan and Amharic, respectively, with `LLaMA-3.1-8B-It`, 1.5 MetricX in Amharic with `LLaMA-3.1-70B-It`, and 1.6 MetricX points in Samoan with `LLaMA-3.1-8B-It`. Small values of k tend to result in smaller gains, and we attribute this to the fact that despite small sentences (phrases) being easier to translate for LLMs, using in-context examples helps them do it in an even better way, particularly into languages with non-Latin scripts.

Second, we replace CompTra's few-shot MT step with a direct translation from `NLLB-200-distilled-600M` (Costa-jussà et al., 2022), a supervised translation model. The LLM draws inspiration from (i.e., combines) the phrases' translations provided by NLLB to produce the final translation. It acts as a merger, and we call this approach NLLB+CompTra. We compare it to few-shot MT with BM25 and CompTra and report the results in Table 1. NLLB+CompTra comes close or outperforms NLLB in all scenarios, proving that the translation of the phrases has a big impact on CompTra's results. NLLB+CompTra outperforming NLLB when the LLM few-shot

	Amharic	Burmese	Fijian	Khmer	Lao	Samoan	Sinhala	Tsonga	Turkmen	Uyghur
LLaMA-3.1-70B-It										
5-shot BM25	13.02	4.26	15.23	4.92	11.02	9.14	3.64	15.88	5.66	4.37
CompTra	**11.95**	**3.64**	**14.94**	4.75	10.54	8.89	**3.28**	**15.38**	5.14	4.30
CompTra with **Words**	17.26	6.26	19.30	7.08	14.01	12.50	5.59	19.49	8.33	7.49
CompTra with **Structure**	16.65	7.64	17.66	6.93	13.38	11.96	6.11	17.67	9.05	8.41
CompTra with **Repeat**	12.96	4.09	15.25	4.90	10.87	9.12	3.53	15.85	5.59	4.30
CompTra with **Paraphrase**	12.86	6.45	15.05	**4.57**	**10.40**	**8.40**	**3.25**	15.58	**4.64**	**4.03**

Table 4: Ablation study on the impact of the decomposition on CompTra (MetricX scores).

	Amharic	Khmer	Lingala	Luganda	Tamil
IN-DOMAIN EVALUATION					
LLaMA-3.1-8B-Instruct					
5-shot BM25	**21.51**	9.79	18.38	19.63	**3.60**
CompTra [with BM25]	21.73	**9.25**	**18.26**	**19.24**	3.77
LLaMA-3.1-70B-Instruct					
5-shot BM25	12.83	4.54	14.65	12.81	2.09
CompTra [with BM25]	**11.97**	**3.73**	**14.51**	**12.76**	**1.52**
OUT-OF-DOMAIN EVALUATION					
LLaMA-3.1-8B-Instruct					
5-shot BM25	**23.52**	13.42	20.45	22.38	**4.92**
CompTra [with BM25]	**23.52**	**13.10**	**19.98**	**21.97**	5.26
LLaMA-3.1-70B-Instruct					
5-shot BM25	14.96	5.70	16.86	**16.07**	2.32
CompTra [with BM25]	**14.29**	**5.13**	**16.69**	**16.03**	**1.81**

Table 5: In-domain and out-of-domain MetricX scores.

performance is better than NLLB's (e.g. Burmese, Khmer etc.) is intuitive. However, it also happens when it is not the case (Lao, Samoan), suggesting that using a strong translator to translate each subpart of a sentence and combining them with a strong "merger" can surpass directly using the translator on the whole sentence. NLLB+CompTra and CompTra both outperform CoD by a large margin, proving that word-level information does not work as well as phrases with context.

What happens when we change the retriever? In all our experiments, CompTra uses BM25 as the retriever. Here, we consider two different retrievers: SONAR and LCS (Longest Common Subsequence). LCS retrieval is based on the longest common subsequence between the query and the candidates after transforming them into space-separated elements. In Table 2 we can see that BM25 is the best retriever for few-shot MT. This superiority is preserved when each retriever is used within the CompTra framework. Ultimately, BM25 is a simple, fast and performing choice.

What happens when we change the decomposition algorithm? The decomposition step uses ICL and MinWikiSplit samples to break sentences into simple propositions, balancing between word-level and sentence-level for context-rich yet manageable phrases for accurate translation. We investigate four strategies: *Words*, *Repeat*, *Paraphrase* and *Structure*. *Words* decomposes a sentence into words while ignoring stop words. *Repeat* uses the main sentence as phrases (with $k = 4$). In the *Paraphrase* strategy, the sentence is paraphrased into at least four variations using a new `divide prompt` (See Appendix C.3). Finally, *Structure* divides the sentence into phrases by heuristically analyzing its dependency tree (see Appendix C.4). In Table 4, we observe that *Words* and *Structure* perform the worst. A common factor in these approaches is that the phrases obtained after decomposition are not full independent sentences, making them difficult to translate. Even with sentence-translation examples in context (few-shot MT), i.e out-of-domain examples, the task remains challenging and the phrase-translation pairs obtained hurt the main MT task. *Repeat* shows marginal improvement over few-shot, indicating that phrase similarity to the main sentence matters, but repetition offers little benefit. *Paraphrase* supports this observation, outperforming *Repeat* and occasionally even slightly surpassing CompTra's native form. It is worth noting that *Paraphrase* uses more phrases on average (4.9) than CompTra's native form (3.2) and is comparatively slower due to more tokens to generate and longer prompts.

6 Conclusion

We introduced a simple yet effective strategy, which we refer to as *Compositional Translation* to improve the MT capabilities of LLMs. Through experiments on three MT benchmarks covering 15 different low-resource directions, we find that it outperforms the strong few-shot MT baseline with similarity search and several strong, existing strategies. It also enables smaller-scale LLMs to elicit better translation capabilities in in-domain and out-of-domain scenarios. Applying compositionality to perform MT will hopefully inspire further work on reasoning-based approaches to MT.

Limitations

In this paper, we introduce CompTra, a simple yet effective approach for improving LLM translation into low-resource languages by exploiting the compositional nature of machine translation. Since our method relies on the LLM's fluency in the source language for sentence decomposition, it is naturally better suited for translation from high-resource languages such as English. Moreover, CompTra is flexible: it can incorporate external modules for sentence decomposition and other components of the translation pipeline.

Acknowledgments

This work was partly funded by Rachel Bawden and Benoît Sagot's chairs in the PRAIRIE institute, now PRAIRIE-PSAI, funded by the French national agency ANR, respectively as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001 and 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. This work was partly supported by compute credits from a Cohere For AI Research Grant, these grants are designed to support academic partners conducting research with the goal of releasing scientific artifacts and data for good projects.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

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.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco 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.

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.

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. 2022. Efficient training of language models to fill in the middle. *arXiv preprint arXiv:2207.14255*.

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, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, and 374 others. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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 forimproved 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.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. 2023. BLASER: A text-free speech-to-speech translation evaluation metric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9064–9079, Toronto, Canada. Association for Computational Linguistics.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. *Preprint*, arXiv:2306.03856.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Cohere. 2024. Command r+. `https://docs.cohere.com/v2/docs/command-r-plus`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-east asia. *Preprint*, arXiv:2404.03608.

Moussa Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory Conde, Kalo Mory Diané, Chris Piech, and Christopher Manning. 2023. Machine translation for nko: Tools, corpora, and baseline results. In *Proceedings of the Eighth Conference on Machine Translation*, pages 312–343, Singapore. Association for Computational Linguistics.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, 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.

Maxim Enis and Mark Hopkins. 2024. From llm to nmt: Advancing low-resource machine translation with claude. *Preprint*, arXiv:2404.13813.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Tear: Improving llm-based

machine translation with systematic self-refinement. *Preprint*, arXiv:2402.16379.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*.

Matthieu Futeral, Armel Zebaze, Pedro Ortiz Suarez, Julien Abadji, Rémi Lacroix, Cordelia Schmid, Rachel Bawden, and Benoît Sagot. 2024. moscar: A large-scale multilingual and multimodal document-level corpus. *Preprint*, arXiv:2406.08707.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1116 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 178 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *Preprint*, arXiv:2302.07856.

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.

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.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, and 3 others. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

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.

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*.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8186–8213, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024. Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA. Association for Computational Linguistics.

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring. *Preprint*, arXiv:2407.03618.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

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, Zhiquan 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.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Christina Niklaus, André Freitas, and Siegfried Handschuh. 2019. MinWikiSplit: A sentence splitting corpus with minimal propositions. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 118–123, Tokyo, Japan. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting 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.

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.

Ratish Puduppully, Anoop Kunchukuttan, Raj Dabre, Ai Ti Aw, and Nancy Chen. 2023. DecoMT: Decomposed prompting for machine translation between related languages using large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4586–4602, Singapore. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Ricardo Rei, Nuno M Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José GC de Souza, and André FT Martins. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. *arXiv preprint arXiv:2309.11925*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. Gaithersburg, MD: NIST.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *The Twelfth International Conference on Learning Representations*.

Davide Turcato and Fred Popowich. 2001. What is example-based machine translation? In *Workshop on Example-Based machine Translation*, Santiago de Compostela, Spain.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George F. Foster. 2023. Prompting palm for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15406–15427. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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.

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*.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2024b. X-alma: Plug & play modules and adaptive rejection for quality translation at scale. *Preprint*, arXiv:2410.03115.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024c. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Forty-first International Conference on Machine Learning*.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024d. 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.

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*.

Armel Zebaze, Benoît Sagot, and Rachel Bawden. 2024a. In-Context Example Selection via Similarity Search Improves Low-Resource Machine Translation. *Preprint*, arXiv:2408.00397.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2024b. 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.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023b. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

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.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A Additional Experiments

MT for Nko. Doumbouya et al. (2023) extended FLORES-200 to incorporate Nko, a language spoken across multiple West African countries. We assess the ability of LLMs to use the Nko writing system, which is significantly different from the other languages we evaluate, and for which neural-based evaluation is still not the standard. Similar to FLORES-200, the selection pool contains 997 sentence pairs and the test set 1012 pairs.
Since SOTA models no longer publicly disclose the content of their training datasets, we cannot rule out the possibility that popular benchmarks might be included in these, as reported by Enis and Hopkins (2024) with Claude 3 Opus and FLORES-200. With Nko, there is a lower risk of such a contamination, allowing for a test of CompTra in a very-low resource scenario.
We evaluate the generations with the n-gram matching metrics BLEU and chrF++ following Doumbouya et al. (2023). In Table 6, we observe that CompTra works well on Nko with gains of up to 4.5 BLEU and 8 chrF++. Usually the few-shot translations contain many repeating tokens, and this issue is alleviated with the use of CompTra.

Ensembling CompTra outperforms few-shot MT with examples retrieved via-similarity search. In this section, we propose to do an ensembling of both approaches with the help of BLASER 2.0 QE (Chen et al., 2023) to account for the strengths of both approaches. Given the 2 candidate translations we choose the one with the highest quality estimation score with respect to the source sentence as the final translation. We compare this ensembling approach against each individual approach and report the results in Table 7. We observe that the ensembling strategy consistently outperforms both of its individual components across all directions considered. This indicates that, while CompTra performs better than standalone few-shot MT, their outputs differ, allowing them to complement and enhance each other.

Non-English-centric directions. We have evaluated LLMs on their ability to generate the translation of english sentences into low-resource languages. In this section, we probe them to translate from French instead. All the prompts follow the same structure as in English, with the `divide prompt` using the same sentences translated into French via Google Translate[9]. We report the results in Table 8. Translating from French is more difficult than from English as indicated by the scores. Overall, CompTra maintains its advantage over few-shot MT, though the performance gap narrows, with a few instances where few-shot MT outperforms CompTra. We attribute this to multiple factors including the quality of the `divide prompt` and the intrinsic abilities of the LLMs in French.

High-resource languages as targets. We evaluate compositional translation when translating from English to five high-resource languages: French (fra), German (deu), Spanish (spa), Portuguese (por) and Japanese (jap). As observed in Table 9, CompTra fails to outperform few-shot MT. We

[9] https://translate.google.com/

Method	LLaMA 3.1 70B It		Gemma 2 27B It		Command-R+	
	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++
5-shot BM25	**8.80**	19.38	10.8	16.49	2.87	6.88
CompTra (Ours)	8.06	**22.16**	**15.53**	**23.04**	**8.59**	**14.55**

Table 6: Full quantitative BLEU and chrF++ results for English→Nko on FLORES-200 Nko's split derived by Doumbouya et al. (2023).

	Amharic	Burmese	Fijian	Khmer	Lao	Samoan	Sinhala	Tsonga	Turkmen	Uyghur
LLaMA-3.1-8B-Instruct										
5-shot BM25	23.40	14.27	21.74	12.63	22.81	19.80	13.79	23.02	14.72	14.01
CompTra (Ours)	23.05	14.29	20.93	12.02	22.42	18.25	13.23	22.75	14.39	15.00
Ensemble	**22.65**	**12.29**	**20.58**	**10.70**	**21.72**	**17.67**	**11.60**	**22.27**	**12.41**	**12.80**
LLaMA-3.1-70B-Instruct										
5-shot BM25	13.02	4.26	15.23	4.92	11.02	9.14	3.64	15.88	5.66	4.37
CompTra (Ours)	11.95	3.64	14.94	4.75	10.54	8.89	3.28	15.38	5.14	4.30
Ensemble	**10.93**	**3.35**	**14.01**	**4.35**	**9.62**	**8.12**	**2.93**	**14.57**	**4.60**	**3.79**

Table 7: Comparison between the ensembling strategy and each of its components (MetricX scores).

observe that zero-shot and few-shot approaches consistently perform best, typically with only a slight difference in performance between them. This partially explains the failure of CompTra, where self-generated in-context demonstrations fail to contribute meaningfully to the MT task and, in some cases, even hinder performance — similar to those retrieved via similarity search. Additionally, we observed that smaller LMs (such as `LLaMA-3.1-8B-It`) occasionally struggle to follow complex instructions included in pipelines like SBYS and TEaR, leading to poor performance. While CompTra avoids these issues due to its simplicity, it still fails to improve translation from English to other high-resource languages.

Reference-free Evaluation While reference-based evaluation metrics are highly correlated with human judgment, they suffer from a reference bias which advantages the translation with a similar style to the reference (Freitag et al., 2020). In Table 15 we observed that CompTra consistently outperforms 5-shot BM25 according to reference-based metrics, now we evaluate if it still holds when using the reference-free MetricX (`MetricX-23-QE-XXL`). In Table 10 we observe that CompTra is still the best strategy, performing better the other across most directions. With COMETKIWI-QE (`wmt23-cometkiwi-da-xxl`; Rei et al., 2023), the conclusion is globally the same but we note a few directions where there is a disagreement with MetricX (Samoan, Tsonga).

Moreover, we use `MetricX-23-QE-XXL` to compare how well LLMs translate the phrases compared to the main sentences. As reported in Table 11, phrases are translated more accurately, confirming CompTra's core hypothesis. Heuristically, we observed that a larger quality gap between phrase translations and main sentence translations correlates with better CompTra performance. However, with the **Paraphrase** strategy, we observe that `LLaMA-3.1-70B-It` better translate a self-generated paraphrase a sentence than the sentence itself but not as well as the short phrases obtained with the native `divide prompt`. Indeed, how good phrases are translated matters but the similarity the semantic similarity between sentence and the phrases seems to be more important for the success of CompTra.

Zero-shot Evaluation Throughout the paper, we have applied CompTra in scenarios where we leveraged the availability of a selection pool. In this section, we apply CompTra in zero-shot i.e. the phrases are translated in zero-shot and the couples pair-translation obtained serve to translate the main sentence in few-shot. In Table 12 we can observe that CompTra performance is not guaranteed. The gains are small with `LLaMA-3.1-70B-It` and `Command-R+` and CompTra fails to improve over zero-shot MT with `Gemma-2-27B-It`.

Impact of the phrase translation quality When comparing the `MetricX-23-QE-XXL` scores across retrievers in Table 13, we observe the same pattern as in Tables 1 and 2. Phrase translation quality is

	Amharic	Burmese	Fijian	Khmer	Lao	Samoan	Sinhala	Tsonga	Turkmen	Uyghur
LLaMA-3.1-8B-Instruct										
5-shot BM25	23.77	**16.22**	22.48	**13.13**	23.46	21.12	**15.58**	23.39	17.12	**16.14**
CompTra (Ours)	**23.74**	17.41	**21.96**	13.85	**23.10**	**20.10**	15.80	**22.97**	**17.09**	17.24
LLaMA-3.1-70B-Instruct										
5-shot BM25	13.95	4.76	16.84	**4.96**	11.89	**9.84**	3.78	**17.14**	6.15	**4.82**
CompTra (Ours)	**13.56**	**4.42**	**16.68**	5.15	**11.40**	10.10	**3.51**	**17.09**	**6.07**	5.19

Table 8: Full MetricX results for ten French→X directions from FLORES-200.

	French	German	Japanese	Portuguese	Spanish
LLaMA 3.1 8B Instruct					
Zero-shot	1.49	1.09	1.39	1.38	1.41
SBYS	13.43	12.54	8.17	11.79	12.32
TEaR	8.20	10.65	12.70	11.21	9.21
5-shot BM25	**1.41**	**1.04**	**1.30**	**1.35**	**1.30**
CompTra (Ours)	1.65	1.20	1.75	1.56	1.52
LLaMA 3.1 70B Instruct					
Zero-shot	**1.03**	**0.68**	1.14	**0.99**	1.17
SBYS	1.06	1.66	10.61	2.09	1.64
TEaR	1.19	0.86	0.97	1.11	1.17
5-shot BM25	**1.02**	**0.69**	**0.78**	1.04	**1.06**
CompTra (Ours)	1.23	0.85	0.97	1.20	1.21

Table 9: Full MetricX results for five English→X high-resource directions from FLORES-200.

consistently higher with BM25 than with LCS or SONAR, which also leads to CompTra performing best when paired with BM25. Furthermore, in cases where NLLB translates phrases better than 5-shot BM25, we observe that NLLB + CompTra outperforms CompTra using BM25 alone, as shown in Table 1. A notable exception is Sinhala: although NLLB yields lower-quality phrase translations than 5-shot BM25, NLLB + CompTra still achieves the best overall results. We hypothesize that in this case, the LLM has stronger internal representations for Sinhala, enabling it to rely less on NLLB outputs and instead leverage its own knowledge to enhance the final translation.

Inference cost of CompTra Given a sentence to translate:

- **CompTra** requires one inference call to decompose the sentence into a certain number of phrases N, N parallel calls to translate these phrases, and one final call to perform N-shot translate the full sentence.
- **TEaR** involves three steps: a first inference call to few-shot translate the sentence, a second to produce MQM annotations, and a third to refine the translation using the annotations.
- **SBYS** uses four inference calls for: pre-drafting research, drafting, refinement, and proofreading.
- **MAPS** requires 3 calls to generate the topic of the sentence, relevant keywords and similar sentences along with the translations. It then performs three more calls to translate using each of these aspects, plus a final call for zero-shot translation. The best of the four candidates is selected using an external QE score.

Chain-of-thought approaches typically require two inference calls: one for generating the reasoning path and one for extracting the final answer, as in Kojima et al. (2022). Similarly, refinement-based methods use two calls per sentence: draft generation followed by refinement. The number of inference calls of CompTra is comparable that of existing strategies such as TEaR, MAPS and SBYS. We also measured the time (in minutes) required to translate the first 100 FLORES sentences in Amharic using `LLaMA-3.1-70B-It` across all techniques, as reported in Table 14. CoT methods are among the slowest due to the lengthy reasoning chains they generate, whereas other approaches are more direct. As expected, zero-shot MT is the fastest, and CompTra ranks among the quickest non-baseline methods.

Methods	Amharic		Burmese		Fijian		Khmer		Lao	
	COMET	MetricX	COMET	MetricX	COMET	MetricX	COMET	MetricX	COMET	MetricX
Zero-shot	46.66	11.97	77.51	3.08	20.72	12.41	77.45	3.71	45.82	10.69
Zero-shot + CoT	29.48	16.59	55.04	7.94	17.70	12.18	69.22	5.32	27.24	16.17
Zero-shot + Refine	48.27	11.68	78.56	2.57	19.95	12.45	78.30	3.44	46.14	10.94
SBYS	48.43	10.86	78.00	2.91	**21.43**	10.74	79.60	3.05	47.10	9.91
MAPS	52.97	9.92	79.55	2.44	18.26	13.86	77.38	3.75	49.58	9.78
TEaR	53.58	9.36	75.51	4.18	20.60	9.77	78.98	3.33	55.66	7.95
5-shot BM25	55.51	8.77	80.16	2.46	21.19	8.39	79.79	3.15	58.55	7.07
+ CoT	45.25	11.19	71.22	4.23	19.72	8.57	76.96	3.53	47.00	9.66
+ Refine	54.95	9.16	**80.94**	1.98	20.37	8.93	**80.38**	**3.03**	57.31	7.58
CompTra (Ours)	**58.76**	**7.58**	**80.94**	**1.93**	**21.38**	**7.74**	80.13	3.07	**59.64**	**6.70**

Methods	Samoan		Sinhala		Tsonga		Turkmen		Uyghur	
	COMET	MetricX	COMET	MetricX	COMET	MetricX	COMET	MetricX	COMET	MetricX
Zero-shot	14.68	7.97	75.63	2.72	23.80	10.60	26.49	5.55	69.57	2.41
Zero-shot + CoT	12.56	9.14	60.51	4.35	21.44	11.83	23.00	6.70	58.29	4.28
Zero-shot + Refine	14.42	7.95	76.12	2.53	23.96	10.64	27.48	4.71	72.92	2.06
SBYS	13.64	7.45	77.73	2.19	22.72	10.41	27.15	4.70	73.40	2.01
MAPS	**15.50**	9.05	78.06	2.21	**25.83**	10.72	26.24	6.05	74.81	2.44
TEaR	15.06	6.69	75.15	3.02	21.74	8.70	28.42	3.64	77.19	2.31
5-shot BM25	14.60	6.22	78.35	2.33	22.39	8.22	28.47	3.36	**78.92**	1.63
+ CoT	13.79	6.66	75.98	2.29	21.62	8.71	27.04	3.75	72.45	2.28
+ Refine	14.40	6.24	78.74	2.24	22.29	8.21	**28.88**	**2.96**	79.55	**1.53**
CompTra (Ours)	14.93	**5.82**	**79.39**	**1.92**	22.60	**7.73**	27.81	**3.01**	78.16	1.62

Table 10: Full COMETKIWI-QE and MetricX-QE results for ten English → X translation directions from FLORES-200 (Goyal et al., 2022; Costa-jussà et al., 2022). We compare CompTra to CoT (Kojima et al., 2022), MAPS (He et al., 2024), SBYS (Briakou et al., 2024) and TEaR (Feng et al., 2024).

	Amharic	Burmese	Fijian	Khmer	Lao	Samoan	Sinhala	Tsonga	Turkmen	Uyghur
LLaMA-3.1-70B-It										
5-shot BM25	8.77	2.46	8.39	3.15	7.07	6.22	2.33	8.22	3.36	1.63
Phrases	3.89	0.88	5.47	1.47	3.34	2.95	0.88	4.97	2.02	1.08
CompTra	7.58	1.93	7.74	3.07	6.70	5.82	1.92	7.73	3.01	1.62
Paraphrase's phrases	8.45	2.15	8.22	2.66	6.34	5.82	1.83	8.20	2.95	1.41
CompTra with **Paraphrase**	8.41	4.94	7.72	2.92	6.56	5.46	1.91	7.63	2.52	1.34
Gemma-2-27B-It										
Phrases	4.65	1.99	7.56	2.11	3.04	5.48	1.84	5.45	1.95	2.18
CompTra	8.19	3.86	10.63	4.23	5.80	9.88	3.68	8.04	2.67	3.32
Command-R+										
Phrases	14.09	4.79	8.09	8.38	7.15	8.87	2.53	10.13	1.77	1.79
CompTra	19.88	8.39	10.74	12.28	12.42	13.85	4.81	15.14	2.45	2.91

Table 11: Full MetricX-QE results for ten English→ X directions. We compare how accurately phrases are translated compared to main sentences.

B Additional Results

B.1 XCOMET and MetricX results on FLORES-200

We report the XCOMET and MetricX of CompTra and baselines on FLORES-200 in Table 15.

B.2 XCOMET and MetricX results against existing approaches

We report the XCOMET and MetricX of CompTra and existing approaches on FLORES-200 in Table 16.

B.3 BLEU and chrF++ results on FLORES-200

We report the BLEU and chrF++ scores of CompTra on FLORES-200, alongside baseline results, in Table 17. The results show the same pattern as the XCOMET and MetricX results shown in the main part of the paper. CompTra outperforms few-shot with SONAR and BM25 in all scenarios. When it comes to LRLs, few-shot MT with example selection via similarity search should be the standard as it always outperforms zero-shot MT and has been proven to perform better than random selection (Zebaze et al., 2024a).

Moreover, as reported in Table 18, CompTra has the interesting property of improving both neural-based and string-matching metrics as opposed to existing strategies. The superiority of CompTra is observed across four distinct metrics (XCOMET, MetricX, BLEU and chrF++), each with unique properties, highlighting its robustness.

	Amharic	Burmese	Fijian	Khmer	Lao	Samoan	Sinhala	Tsonga	Turkmen	Uyghur
LLaMA-3.1-70B-It										
Zero-shot	16.95	5.18	19.48	5.69	16.67	11.51	4.27	**19.82**	8.77	8.36
CompTra (Zero-shot)	17.17	4.97	20.04	5.88	17.44	12.38	4.11	20.93	8.33	6.92
Ensemble Zero-shot	**15.54**	**4.11**	**18.87**	**5.15**	**15.93**	**10.98**	**3.51**	**19.82**	**7.20**	**6.15**
Gemma-2-27B-It										
Zero-shot	15.49	**9.05**	20.39	9.15	**13.40**	17.89	7.48	20.02	6.07	12.90
CompTra (Ours)	15.86	11.61	20.91	10.13	15.09	18.91	8.45	20.90	6.44	14.50
Ensemble Zero-shot	**14.44**	**9.08**	**20.04**	**8.86**	13.59	**17.52**	**6.89**	**19.65**	**5.38**	**12.30**
Command-R+										
Zero-shot	24.46	19.63	22.95	19.44	21.71	23.21	10.62	24.36	6.04	14.00
CompTra (Ours)	24.41	19.12	22.69	19.21	21.67	22.97	10.33	24.41	6.06	13.90
Ensemble Zero-shot	**24.26**	**17.84**	**22.14**	**17.84**	**20.81**	**22.55**	**8.71**	**24.24**	**5.01**	**11.71**

Table 12: CompTra's performance in zero-shot (MetricX scores).

	Amharic	Burmese	Fijian	Khmer	Lao	Samoan	Sinhala	Tsonga	Turkmen	Uyghur
LLaMA-3.1-70B-It										
5-shot LCS	4.97	1.18	5.89	1.73	4.51	3.60	1.08	5.33	2.19	1.14
5-shot SONAR	4.37	0.95	5.58	1.52	3.97	3.29	**0.90**	5.12	2.06	1.14
5-shot BM25	3.89	**0.88**	5.47	**1.47**	3.34	2.95	**0.89**	4.97	**2.02**	**1.08**
NLLB (zero-shot)	**2.79**	1.28	**4.99**	3.91	**1.88**	2.17	1.28	**3.52**	2.57	2.00

Table 13: Full MetricX-QE results for ten English→ X directions. We compare how accurately phrases are translated across different retrievers.

B.4 XCOMET and MetricX results on NTREX 128 and TICO-19

We report the results obtained by CompTra on NTREX 128 and TICO-19 in Tables 19 and 20 respectively.

B.5 BLEU and chrF++ results on NTREX 128 and TICO-19

We present results with BLEU and chrF++ scores of CompTra against baselines on NTREX 128 in Table 21 and TICO-19 in Table 22. The results are the same as in Table 19 and Table 20 where CompTra outperforms few-shot MT with BM25.

B.6 BLEU and chrF++ results of the Out-of-domain evaluation

We present results with BLEU and chrF++ scores of CompTra against baselines in the out-of-domain setup in Table 23. The results are the same as in Table 5 where CompTra outperforms few-shot MT with BM25.

B.7 BLEU and chrF++ results on FLORES 200 with small LMs

We reported that CompTra works very well with smaller LMs in Table 3 by reporting the MetricX scores. In Table 24, we show that the performance gains provided by CompTra are also observable in terms of BLEU and chrF++. Moreover, we compare SBYS, TEaR, MAPS and 5-shot BM25 + Self-refine to CompTra using `LLaMA-3.1-8B-It`, `Gemma-2-9B-It` and `Command-R` and report the results in Table 25. CompTra ends up being the best approach at this scale too. The models sometime struggle to directly refine their answers, leading the performances of 5-shot BM25 + Self-refine to be worse or equal to 5-shot BM25. SBYS does not work well with `LLaMA-3.1-8B-It` but give good results with `Gemma-2-9B-It` (Similar to the strong results they achieved with Gemini; Gemini Team et al., 2024), outperforming CompTra in some scenarios. With `Command-R`, SBYS does not fail as it does with `LLaMA-3.1-8B-It` but it remains worse than CompTra in most scenarios.

B.8 How often are CompTra's phrases translated in an incorrect language on FLORES 200?

CompTra performs language identification on the translation of each phrase obtained after decomposition. This step is necessary to prevent the noise from a different language to be inadvertently incorporated into the translation of the main sentence. It is not required for the few-shot baseline, as in-context demonstrations are directly retrieved from a curated set of high-quality trans-

Zero-shot	Zero-shot+CoT	Zero-shot+Refine	5-shot	5-shot+CoT	5-shot+Refine	TEaR	MAPS	SBYS	CompTra
7.0	60.1	25.4	14.4	60.4	25.1	36.6	85.5	41.1	34.9

Table 14: Time (in minutes) required to translate the first 100 FLORES sentences in Amharic using `LLaMA-3.1-70B-It`.

Methods	Amharic		Burmese		Fijian		Khmer		Lao	
	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX
LLaMA-3.1-70B-Instruct										
Zero-shot	31.25	16.95	57.73	5.18	20.44	19.48	60.76	5.69	32.63	16.67
5-shot SONAR	37.51	13.61	60.91	4.50	21.84	15.52	64.00	4.92	42.05	11.55
5-shot BM25	39.55	13.02	**62.27**	4.26	22.18	15.23	**64.46**	4.92	45.38	11.02
CompTra (Ours)	**41.32**	**11.95**	60.84	**3.64**	**22.39**	**14.94**	64.22	**4.75**	**47.17**	**10.54**
Gemma-2-27B-It										
Zero-shot	32.79	15.49	40.07	9.05	19.91	20.39	45.37	9.15	38.80	13.40
5-shot SONAR	37.07	13.69	47.38	7.02	21.00	18.20	53.05	7.14	47.17	10.58
5-shot BM25	38.09	13.23	**48.62**	**6.80**	20.96	18.17	54.00	**7.02**	48.03	10.50
CompTra (Ours)	**40.10**	**12.64**	47.98	7.23	**21.55**	**16.86**	**55.02**	**7.05**	**51.02**	**9.88**
Command-R+										
Zero-shot	16.32	24.46	24.39	19.63	18.63	22.95	23.39	19.44	19.81	21.71
5-shot SONAR	18.06	23.45	29.60	15.12	20.13	19.52	27.91	18.45	25.48	18.46
5-shot BM25	17.96	23.24	32.25	14.76	20.30	19.19	28.37	18.40	26.74	18.03
CompTra (Ours)	**19.33**	**22.54**	**35.60**	**12.34**	**21.16**	**16.85**	**31.76**	**15.53**	**29.64**	**16.52**

Methods	Samoan		Sinhala		Tsonga		Turkmen		Uyghur	
	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX
LLaMA-3.1-70B-Instruct										
Zero-shot	23.61	11.51	64.72	4.27	21.91	19.82	23.46	8.77	43.65	8.36
5-shot SONAR	24.42	9.30	67.60	3.60	23.66	16.45	25.14	5.83	58.16	**4.33**
5-shot BM25	24.78	9.14	67.64	3.64	24.46	15.88	24.97	5.66	**59.70**	4.37
CompTra (Ours)	**24.95**	**8.89**	**68.63**	**3.28**	**24.77**	**15.38**	**25.40**	**5.14**	58.32	**4.30**
Gemma-2-27B-It										
Zero-shot	21.38	17.89	48.27	7.48	21.81	20.02	25.26	6.07	29.83	12.90
5-shot SONAR	22.30	15.16	52.30	6.65	23.84	16.39	25.78	4.53	36.34	9.98
5-shot BM25	22.69	14.59	53.68	6.50	24.06	15.97	**26.26**	**4.48**	37.52	9.72
CompTra (Ours)	**23.16**	**13.80**	**54.47**	**6.19**	**24.53**	**15.02**	25.98	4.55	**38.97**	**9.37**
Command-R+										
Zero-shot	18.84	23.21	38.56	10.62	19.01	24.36	25.03	6.04	29.12	14.00
5-shot SONAR	20.30	20.15	45.60	8.65	20.56	22.41	24.93	4.70	39.74	9.46
5-shot BM25	20.75	19.48	47.06	8.11	20.83	22.08	25.11	4.57	42.00	8.54
CompTra (Ours)	**21.78**	**17.31**	**49.29**	**6.99**	**21.89**	**21.59**	**25.26**	**4.07**	**44.14**	**7.64**

Table 15: XCOMET and MetricX scores for 10 English → X directions from FLORES-200 (Goyal et al., 2022; Costa-jussà et al., 2022). Best results (including any results that are not statistically worse) are highlighted in bold.

lation pairs produced by professional translators in the correct languages. As shown in Table 26, with large instruction-tuned language models, the phrase translations are almost always in the correct language. This suggests that CompTra could still perform well without the language identification step, as already demonstrated in Table 6 with Nko.

Methods	Amharic		Burmese		Fijian		Khmer		Lao	
	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX-23
Zero-shot	31.25	16.95	57.73	5.18	20.44	19.48	60.76	5.69	32.63	16.67
Zero-shot + CoT	23.35	22.62	38.61	13.64	19.97	20.55	50.77	8.90	22.98	22.83
Zero-shot + Refine	32.21	16.56	58.92	4.76	20.39	19.56	61.39	5.28	32.98	16.82
SBYS	31.77	16.55	56.79	5.16	20.28	19.27	61.89	4.86	31.58	16.40
MAPS	35.63	14.28	60.94	4.16	20.37	19.15	61.47	5.37	37.69	14.41
TEaR	37.96	13.69	59.89	6.02	21.70	16.32	63.49	5.13	42.83	12.17
5-shot BM25	39.55	13.03	62.27	4.26	22.18	15.23	**64.46**	4.92	45.38	11.02
+ CoT	32.39	18.18	52.07	7.18	22.08	15.97	60.80	5.80	36.78	15.90
+ Refine	38.38	13.46	**62.97**	3.82	21.88	15.96	64.39	**4.76**	43.94	11.59
CompTra (Ours)	**41.32**	**11.95**	60.84	**3.64**	**22.39**	**14.94**	64.22	**4.75**	**47.17**	**10.54**

Methods	Samoan		Sinhala		Tsonga		Turkmen		Uyghur	
	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX-23
Zero-shot	23.61	11.51	64.72	4.27	21.91	19.82	23.46	8.77	43.65	8.36
Zero-shot + CoT	22.58	14.08	49.44	8.47	21.33	21.45	22.83	10.78	38.25	11.53
Zero-shot + Refine	23.49	11.26	66.27	4.11	22.18	19.68	23.65	7.77	47.38	7.60
SBYS	22.54	11.70	65.31	3.61	21.13	20.25	23.83	7.63	47.82	6.28
MAPS	23.12	11.89	68.14	3.41	23.30	18.77	24.44	9.03	51.88	6.20
TEaR	24.45	9.72	65.94	4.38	24.01	16.34	25.15	5.73	57.20	5.06
5-shot BM25	24.78	9.14	67.64	3.64	24.46	15.88	24.97	5.66	59.70	4.37
+ CoT	24.09	10.65	65.64	4.06	24.08	17.22	24.96	6.22	52.46	6.10
+ Refine	24.34	9.26	**69.13**	3.49	24.23	16.15	24.74	5.26	**60.50**	**4.17**
CompTra (Ours)	**24.95**	**8.89**	68.63	**3.28**	**24.77**	**15.38**	**25.40**	**5.14**	58.32	4.30

Table 16: Full XCOMET and MetricX results 10 English→X directions from FLORES-200. We compare CompTra to CoT (Kojima et al., 2022), MAPS (He et al., 2024), SBYS (Briakou et al., 2024) and TEaR (Feng et al., 2024).

Methods	Amharic		Burmese		Fijian		Khmer		Lao	
	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++
LLaMA 3 70B It										
Zero-shot	8.81	16.95	17.41	34.78	7.70	28.05	18.46	30.81	8.30	24.52
5-shot SONAR	11.24	20.25	19.05	36.22	11.85	34.98	20.36	33.43	14.09	31.96
5-shot BM25	11.86	20.92	19.49	36.80	12.19	35.34	20.14	37.53	15.13	32.70
CompTra (Ours)	**12.64**	**22.67**	**19.84**	**38.03**	**12.94**	**38.02**	**20.60**	**33.06**	**16.37**	**33.14**
Gemma 2 27B It										
Zero-shot	7.85	17.24	10.70	28.50	7.39	28.57	11.30	25.13	8.91	25.05
5-shot SONAR	10.19	19.86	13.44	31.72	9.97	31.74	14.36	29.16	15.14	33.75
5-shot BM25	10.52	20.24	13.98	32.08	9.93	31.88	14.57	28.86	15.58	33.73
CompTra (Ours)	**11.67**	**21.46**	**15.11**	**32.99**	**11.63**	**35.72**	**15.61**	**29.52**	**17.20**	**34.62**
Command-R+										
Zero-shot	2.59	8.21	5.32	21.60	4.36	22.45	5.96	19.47	4.15	21.04
5-shot SONAR	4.40	10.69	9.40	27.18	8.38	29.74	8.32	22.18	8.72	27.21
5-shot BM25	4.65	11.17	9.70	27.63	8.94	30.30	8.54	21.77	**9.30**	27.54
CompTra (Ours)	**6.66**	**14.56**	**12.30**	**30.73**	**11.20**	**36.36**	**10.99**	**24.97**	9.10	**28.72**

Methods	Samoan		Sinhala		Tsonga		Turkmen		Uyghur	
	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++
LLaMA 3 70B It										
Zero-shot	16.04	38.25	23.71	36.20	6.34	25.34	13.40	32.69	11.51	25.04
5-shot SONAR	20.93	42.90	25.34	38.03	10.28	31.96	17.99	38.01	19.41	37.05
5-shot BM25	21.47	43.29	25.40	38.02	11.13	33.11	18.56	38.66	20.01	37.53
CompTra (Ours)	**21.59**	**44.53**	**26.20**	**39.67**	**11.55**	**35.01**	**19.69**	**40.64**	**21.11**	**39.00**
Gemma 2 27B It										
Zero-shot	10.63	32.83	14.75	27.41	6.77	27.25	10.17	30.21	6.27	22.62
5-shot SONAR	13.87	36.35	17.96	30.61	10.22	32.34	13.85	35.03	11.07	28.08
5-shot BM25	14.25	36.77	18.57	31.01	10.90	33.13	14.85	35.60	11.90	28.55
CompTra (Ours)	**15.34**	**38.70**	**20.24**	**32.67**	**11.65**	**35.07**	**16.11**	**36.78**	**13.32**	**30.47**
Command-R+										
Zero-shot	5.18	19.90	12.98	26.80	3.19	14.19	13.94	34.17	7.48	23.10
5-shot SONAR	10.96	28.24	16.56	30.31	5.73	21.61	18.64	39.08	13.26	29.93
5-shot BM25	12.19	29.66	17.45	31.02	6.24	22.68	19.69	39.85	14.74	31.57
CompTra (Ours)	**14.67**	**36.37**	**19.64**	**33.71**	**7.84**	**27.80**	**20.72**	**41.09**	**16.59**	**33.77**

Table 17: Full BLEU and chrF++ results for ten English→X directions from FLORES-200 (Goyal et al., 2022; Costa-jussà et al., 2022).

Methods	Amharic		Burmese		Fijian		Khmer		Lao	
	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++
Zero-shot	8.81	16.95	17.41	34.78	7.70	28.05	18.46	30.81	8.30	24.52
Zero-shot + CoT	6.64	13.60	11.42	29.00	6.48	26.86	15.13	27.98	3.61	16.87
Zero-shot + Refine	8.53	16.65	16.91	34.02	7.70	28.41	17.79	29.94	7.60	22.97
SBYS	8.76	17.44	16.43	34.12	7.92	30.04	17.62	29.95	8.18	25.31
MAPS	9.47	17.95	17.42	34.57	6.25	23.41	18.48	31.66	9.21	25.37
TEaR	11.19	20.04	17.93	34.41	11.12	34.15	19.56	32.30	13.62	30.84
5-shot BM25	11.86	20.92	19.49	36.80	12.19	35.34	20.14	**33.15**	15.13	32.70
+ CoT	10.56	19.13	17.06	34.26	10.67	34.47	18.47	31.29	11.29	27.81
+ Refine	11.05	19.82	18.66	35.85	11.51	34.76	18.80	31.10	13.36	29.65
CompTra (Ours)	**12.64**	**22.67**	**19.84**	**38.03**	**12.94**	**38.02**	**20.60**	33.06	**16.37**	**33.14**

Methods	Samoan		Sinhala		Tsonga		Turkmen		Uyghur	
	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++
Zero-shot	16.04	38.25	23.71	36.20	6.34	25.34	13.40	32.69	11.51	25.04
Zero-shot + CoT	13.66	35.55	19.53	32.36	5.26	24.05	11.68	31.12	12.14	27.53
Zero-shot + Refine	15.82	37.98	22.79	35.19	6.20	25.26	13.64	33.14	11.86	25.48
SBYS	14.85	37.54	22.91	36.00	6.14	26.16	13.18	33.02	14.80	31.76
MAPS	15.02	36.20	23.16	35.45	5.89	23.37	11.75	29.47	14.28	29.41
TEaR	19.71	42.02	24.67	36.76	9.88	31.74	16.53	36.97	18.60	36.18
5-shot BM25	21.47	43.29	25.40	38.02	11.13	33.11	18.56	38.66	20.01	37.53
+ CoT	17.72	40.06	24.68	37.37	9.51	31.50	17.22	37.30	19.38	36.25
+ Refine	19.03	41.45	24.19	36.44	10.24	32.18	17.57	37.51	19.08	36.17
CompTra (Ours)	**21.59**	**44.53**	**26.20**	**39.67**	**11.55**	**35.01**	**19.69**	**40.64**	**21.11**	**39.00**

Table 18: Full BLEU and chrF++ results for ten English→X directions from FLORES-200 (Goyal et al., 2022; Costa-jussà et al., 2022). We compare CompTra to CoT (Kojima et al., 2022), MAPS (He et al., 2024), SBYS (Briakou et al., 2024) and TEaR (Feng et al., 2024).

Methods	Amharic		Fijian		Shona		Somali		Tswana	
	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX-23
LLaMA 3.1 70B It										
5-shot BM25	30.62	16.14	21.61	15.80	24.43	15.52	40.39	8.99	26.24	12.40
CompTra (Ours)	**31.61**	**15.46**	**21.75**	**15.38**	**24.69**	**15.18**	**40.70**	**8.85**	**26.73**	**12.15**
Gemma 2 27B It										
5-shot BM25	29.67	16.01	20.54	18.34	25.27	**10.50**	41.89	8.25	26.49	11.49
CompTra (Ours)	**30.46**	**15.40**	**21.04**	**17.35**	**25.63**	**10.48**	**42.34**	**8.11**	**26.69**	**11.42**
Command-R+										
5-shot BM25	18.33	23.05	19.96	18.77	21.42	19.26	25.86	16.78	22.55	18.94
CompTra (Ours)	17.80	22.98	**20.66**	**17.42**	**22.21**	**18.43**	**26.97**	**15.53**	**23.95**	**17.96**

Table 19: XCOMET and MetricX results for 5 English→X directions from NTREX 128 (Federmann et al., 2022).

Methods	Amharic		Khmer		Lingala		Luganda		Tamil	
	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX	XCOMET	MetricX-23
LLaMA 3.1 70B It										
5-shot BM25	39.71	12.71	67.55	4.51	**23.71**	14.69	26.70	12.77	68.00	2.06
CompTra (Ours)	**40.40**	**11.83**	**68.96**	**3.72**	23.65	**14.53**	26.66	**12.68**	**68.46**	**1.50**
Gemma 2 27B It										
5-shot BM25	40.80	11.99	60.42	5.71	24.03	**13.90**	26.45	14.99	**68.47**	**1.50**
CompTra (Ours)	**41.84**	**11.30**	**62.25**	**5.38**	**24.21**	13.98	**26.67**	**13.90**	67.41	1.62
Command-R+										
5-shot BM25	22.45	21.40	38.65	13.69	**22.68**	16.94	23.29	20.24	**66.86**	2.15
CompTra (Ours)	**23.03**	**21.08**	**40.17**	**12.36**	**22.70**	**16.22**	**23.93**	**19.61**	66.34	**1.71**

Table 20: XCOMET and MetricX results for 5 English→X directions from TICO-19 (Anastasopoulos et al., 2020).

Methods	Amharic		Fijian		Shona		Somali		Tswana	
	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++
LLaMA 3.1 70B It										
5-shot BM25	8.11	15.28	12.42	35.59	11.60	31.67	12.76	37.12	18.67	39.87
CompTra (Ours)	**9.13**	**16.71**	**13.42**	**38.30**	**11.87**	**33.92**	**13.21**	**38.40**	**20.07**	**41.61**
Gemma 2 27B It										
5-shot BM25	6.99	15.16	10.25	32.34	12.78	35.33	12.65	37.11	18.66	40.83
CompTra (Ours)	**8.33**	**16.55**	**12.48**	**36.53**	**13.55**	**36.46**	**13.27**	**37.69**	**19.84**	**42.19**
Command-R+										
5-shot BM25	3.09	8.82	9.46	30.37	7.50	25.08	8.45	28.14	11.77	29.57
CompTra (Ours)	**4.61**	**11.38**	**11.92**	**36.21**	**9.63**	**29.72**	**9.93**	**32.09**	**14.98**	**35.82**

Table 21: Full BLEU and chrF++ results for ten English→X directions from NTREX 128 (Federmann et al., 2022; Barrault et al., 2019).

Methods	Amharic		Khmer		Lingala		Luganda		Tamil	
	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++
LLaMA 3.1 70B It										
5-shot BM25	11.90	20.90	32.94	44.08	14.47	35.63	15.83	36.00	32.05	50.43
CompTra (Ours)	**13.44**	**23.08**	**34.71**	**45.60**	**15.17**	**39.21**	**16.21**	**37.86**	**33.60**	**52.02**
Gemma 2 27B It										
5-shot BM25	10.95	21.25	26.56	39.96	15.27	37.51	14.00	33.77	27.56	47.53
CompTra (Ours)	**12.50**	**22.81**	**28.67**	**41.40**	**16.11**	**39.91**	**15.77**	**36.52**	**28.62**	**48.21**
Command-R+										
5-shot BM25	5.90	12.98	18.34	31.18	11.64	31.00	8.54	24.56	28.77	47.62
CompTra (Ours)	**8.25**	**16.74**	**20.76**	**34.11**	**14.19**	**36.86**	**11.18**	**30.05**	**29.63**	**48.69**

Table 22: Full BLEU and chrF++ results for 5 English→X directions from TICO-19 (Anastasopoulos et al., 2020).

Methods	Amharic		Khmer		Lingala		Luganda		Tamil	
	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++
LLaMA 3.1 8B It										
5-shot BM25	2.92	9.94	12.81	28.27	7.96	25.14	5.33	20.07	16.83	38.63
CompTra (Ours)	**3.86**	**12.07**	**13.46**	**29.60**	**10.39**	**31.65**	**6.61**	**24.46**	**17.18**	**39.17**
LLaMA 3.1 70B It										
5-shot BM25	9.49	18.17	27.33	39.82	11.06	31.60	9.39	29.26	26.69	47.38
CompTra (Ours)	**10.44**	**20.04**	**27.59**	**40.84**	**12.36**	**36.50**	**9.67**	**31.26**	**27.13**	**48.36**

Table 23: Full BLEU and chrF++ results for 5 English→X directions for the Out-of-domain evaluation.

Methods	Amharic		Burmese		Fijian		Khmer		Lao	
	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++
LLaMA 3.1 8B It										
5-shot BM25	4.27	11.37	10.09	28.09	7.37	25.92	10.44	25.17	4.17	18.84
CompTra (Ours)	**5.76**	**13.73**	**11.62**	**29.88**	**8.60**	**31.32**	**11.68**	**26.30**	**3.53**	**19.74**
Gemma 2 9B It										
5-shot BM25	8.34	17.66	9.69	27.91	8.70	29.26	10.66	25.57	10.77	28.79
CompTra (Ours)	**9.86**	**19.66**	**11.35**	**30.05**	**10.61**	**34.13**	**12.09**	**26.95**	**13.64**	**32.07**
Command-R										
5-shot BM25	2.66	8.85	5.69	22.48	7.83	29.25	5.31	19.21	**4.51**	20.65
CompTra (Ours)	**4.13**	**12.02**	**8.50**	**27.09**	**8.24**	**34.38**	**7.06**	**20.90**	4.47	**22.48**

Methods	Samoan		Sinhala		Tsonga		Turkmen		Uyghur	
	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++	BLEU	chrF++
LLaMA 3.1 8B It										
5-shot BM25	12.01	29.53	12.69	24.67	5.53	21.07	8.93	26.22	10.65	27.20
CompTra (Ours)	**13.95**	**34.46**	**14.84**	**27.71**	**7.43**	**27.53**	**9.80**	**29.38**	**11.90**	**29.15**
Gemma 2 9B It										
5-shot BM25	14.98	34.59	16.04	29.99	6.86	23.70	9.92	29.33	5.83	18.25
CompTra (Ours)	**16.77**	**38.42**	**18.09**	**31.77**	**8.71**	**29.54**	**11.24**	**31.56**	**9.14**	**24.32**
Command-R										
5-shot BM25	8.88	25.48	10.21	23.44	5.92	22.42	13.62	33.88	9.04	24.30
CompTra (Ours)	**11.84**	**33.18**	**13.36**	**27.81**	**6.60**	**27.68**	**14.83**	**35.60**	**10.67**	**28.50**

Table 24: Full quantitative BLEU and chrF++ for ten English→X directions from FLORES-200 (Goyal et al., 2022; Costa-jussà et al., 2022) with small LMs.

	Amharic	Burmese	Fijian	Khmer	Lao	Samoan	Sinhala	Tsonga	Turkmen	Uyghur
LLaMA-3.1-8B-Instruct										
MAPS	24.38	15.49	23.28	13.50	24.53	22.96	13.39	23.94	12.54	16.45
SBYS	24.76	22.14	24.66	16.55	24.92	24.51	23.25	24.67	22.10	22.36
TEaR	24.05	16.98	22.71	14.05	23.51	20.95	16.48	23.64	17.98	17.46
5-shot BM25	23.40	14.27	21.74	12.63	22.81	19.80	13.79	23.02	14.72	**14.01**
+ Refine	23.54	**14.23**	22.34	14.00	23.76	20.77	14.16	23.20	**14.18**	14.46
CompTra	**23.06**	14.29	**20.93**	**12.02**	**22.41**	**18.25**	**13.23**	**22.75**	14.39	15.00
Gemma-2-9B-It										
MAPS	15.42	14.35	21.86	12.40	14.28	20.57	8.98	22.14	5.66	22.38
SBYS	**15.04**	15.30	**18.81**	12.18	13.97	**15.92**	11.01	**18.34**	**5.01**	20.06
TEaR	15.75	13.21	20.54	11.30	14.52	17.14	**8.70**	21.60	8.55	20.96
5-shot BM25	15.99	13.05	20.66	11.92	15.21	17.61	9.13	20.99	8.36	21.07
+ Refine	15.64	13.30	20.93	11.73	15.04	17.71	**8.73**	21.56	5.19	20.89
CompTra	15.66	**12.31**	19.63	**11.23**	**13.67**	**15.93**	8.82	19.82	7.69	**19.19**
Command-R										
MAPS	24.87	22.28	23.85	23.09	23.79	23.58	15.15	24.36	9.61	16.95
SBYS	**23.57**	23.09	22.98	23.67	23.68	22.68	16.78	22.26	7.57	20.43
TEaR	24.57	21.77	21.39	21.73	22.87	21.65	15.94	22.70	7.38	16.53
5-shot BM25	24.38	20.94	21.24	21.64	22.68	21.67	15.50	22.46	7.01	16.36
+ Refine	24.57	21.35	22.30	21.78	23.08	22.78	15.21	23.46	6.60	15.99
CompTra	24.39	**19.33**	**20.59**	**20.48**	**21.88**	**20.82**	**12.91**	**22.16**	**5.95**	**14.99**

Table 25: Full MetricX results for ten English → X translation directions from FLORES-200 with small LMs. We compare CompTra to Self-refine (Chen et al., 2024), MAPS (He et al., 2024), SBYS (Briakou et al., 2024) and TEaR (Feng et al., 2024).

	Amharic	Burmese	Fijian	Khmer	Lao	Samoan	Sinhala	Tsonga	Turkmen	Uyghur
LLaMA-3.1-70B-It	99.8	99.9	99.0	99.8	99.4	98.7	100.0	98.3	98.4	100.0
Command-R+	99.8	99.7	99.5	99.6	99.4	99.1	99.9	98.5	98.5	99.9
Gemma-2-27B-It	99.4	99.5	99.2	99.6	98.4	98.2	98.1	98.5	96.0	99.8
LLaMA-3.1-8B-It	99.6	99.6	98.6	99.6	99.4	98.3	99.8	97.0	97.6	99.6
Command-R	99.8	99.7	99.5	99.9	99.6	99.2	100.0	98.5	98.1	99.8
Gemma-2-9B-It	98.3	99.8	99.0	99.6	99.4	98.8	99.9	98.4	95.0	99.9

Table 26: Percentage of phrases' 5-shot translations written in the correct target language when using CompTra on FLORES-200.

C Implementation details

C.1 General remarks

When translating into LRLs, particularly languages with non-Latin scripts, it is important to generate the right amount of tokens. Current tokenizers tend to require more tokens for non-Latin scripts, thus translating a 100-token English sentence in French can use half as much tokens as doing so in Amharic. This is the reason why we set `max_new_tokens` to 500. However, it comes with the risk of overgeneration (Bawden and Yvon, 2023; Zebaze et al., 2024a). It occurs when translating into LRLs with base models but also with intruction fine-tuned/chat models and usually take the form of repeating n-grams at the end of the generations. This is done by space-separating the generation, identifying a bigram which occurs more than eight times and drop the rest of the sentence after its first occurrence. To assess statistical significance between each pair of strategies, we follow Koehn (2004) and apply paired bootstrap resampling, using 300 samples of 500 sentences each and a significance threshold of $p < 0.05$.

In this paper, a phrase is to be understood as contiguous subpart of a sentence, in the context of phrase-based MT and not in the linguistic sense of a phrase or constituent.

C.2 Models, Datasets and Tools

In Table 27, we list the links to the relevant resources used for experiments.

C.3 Prompts

C.3.1 Translation prompt

Zero-shot

```
Please write a high-quality Amharic
    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.
```

Few-shot

```
Given the following sentence-translation
     pairs written by a professional
    translator:

<Demonstrations>
1. English sentence
"If it becomes commercial, we should
    have it. That is, there's no in-
    principle objection to nuclear
    energy" Mr Costello said.
Amharic translation
<>

2. English sentence
The governor also stated, "Today, we
    learned that some school aged
    children have been identified as
    having had contact with the patient
    ."
Amharic translation
<>

3. English sentence
The commissioner said, "We haven't yet
    agreed on rules of origin and tariff
     con[c]essions, but the framework we
     have is enough to start trading on
    July 1, 2020".
Amharic translation
<>

4. English sentence
Permits are limited to protect the
    canyon, and become available on the
    1st day of the month, four months
    prior to the start month.
Amharic translation
<>

5. English sentence
We have a year-long financial crisis,
    which has had its most acute moment
    in the past two months, and I think
    now the financial markets are
    beginning to recover."
Amharic translation
<>
</Demonstrations>

Please write a high-quality Amharic
    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 make sure to consider the above
    information and provide only the
    translation, nothing more.
```

C.3.2 Divide prompt

Vanilla

```
We would like to derive a list of short
    sentences from long and convoluted
    sentences. For each long sentence,
    you will use punctuation (e.g.,
    comma, semicolon, etc.),
    coordinating conjunctions (e.g., for
    , and, etc.), and subordinating
    conjunctions (e.g., although,
    because) to divide the sentence into
     multiple clauses, which you will
    then use to write simpler sentences.
```

Datasets	
FLORES-200	https://huggingface.co/datasets/facebook/flores
Machine Translation for Nko	https://github.com/common-parallel-corpora/common-parallel-corpora
NTREX	https://github.com/MicrosoftTranslator/NTREX/tree/main
NTREX HF	hhttps://huggingface.co/datasets/mteb/NTREX
TICO-19	https://huggingface.co/datasets/gmnlp/tico19
Models evaluated	
Command-R	command-r-08-2024
Command-R+	command-r-plus-08-2024
Gemma 2 2B It	https://huggingface.co/google/gemma-2-2b-it
Gemma 2 9B It	https://huggingface.co/google/gemma-2-9b-it
Gemma 2 27B It	https://huggingface.co/google/gemma-2-27b-it
LLaMA 3.1 8B It	https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct
LLaMA 3.1 70B It	https://huggingface.co/hugging-quants/Meta-Llama-3.1-70B-Instruct-AWQ-INT4
NLLB-200-distilled-600M	https://huggingface.co/facebook/nllb-200-distilled-600M
Other resources	
MetricX23-XXL	https://huggingface.co/google/metricx-23-xxl-v2p0
XCOMET-XXL	https://huggingface.co/Unbabel/XCOMET-XXL
wmt23-cometkiwi-da-xxl	https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xxl
FastText	https://huggingface.co/facebook/fasttext-language-identification
BM25s	https://github.com/xhluca/bm25s

Table 27: Links to datasets, benchmarks and models.

```
    Ensure that each of the short
    sentences reflects a part of the
    larger sentence. Here are some
    examples.

###

Sentence
The Boolean satisfiability problem is a
    well-researched problem with many
    exemplar solvers available; it is
    very fast, as package solving
    complexity is very low compared to
    other areas where SAT solvers are
    used.

Propositions
    1. The Boolean satisfiability
        problem is a well-researched
        problem.
    2. It has many exemplar solvers are
        available.
    3. It is very fast.
    4. The package solving complexity is
         very low.
    5. This is compared to other areas
        where SAT solvers are used.

###

Sentence
Dore was offered several one-off shows
    in night clubs, and her best album
    was rereleased in 2001.

Propositions
    1. Dore was offered several one-off
        shows in night clubs.
    2. Her best album was rereleased in
        2001.

###

Sentence
Jim briefly transfers to the Stamford
    branch after Pam confirmed her
    commitment to Roy, before corporate
    is forced to merge the Stamford
    branch and staff into the Scranton
    branch.

Propositions
    1. Jim briefly transfers to the
        Stamford branch.
    2. Pam confirmed her commitment to
        Roy.
    3. Corporate is forced to merge the
        Stamford branch and staff.
    4. The merge is into the Scranton
        branch.

###

Sentence
But Jack could not get back to his own
    time, because one of the drug vials
    had broke, and there was only enough
     left in one of the vials to stop
    Whistler.

Propositions
    1. But Jack could not get back to
        his own time.
    2. One of the drug vials had broke.
    3. There was only enough left in one
         of the vials.
    4. This was to stop Whistler.

###

Sentence
However, his nonconformist background
    came to the fore again when he
    became friendly with William Durning
     around 1817, having rented a
    cottage from another member of the
```

```
    Durning family, and on 1 September
    1820 he married William's daughter,
    Emma.

Propositions
    1. However, his nonconformist
        background came to the fore
        again.
    2. He became friendly with William
        Durning around 1817.
    3. He rented a cottage from another
        member of the Durning family.
    4. He married William's daughter.
    5. The marriage was on 1 September
        1820.

###

Sentence
Mallzee was founded in December 2012 by
    Cally Russell and is based in
    Edinburgh.

Propositions
    1. Mallzee was founded in December
        2012 by Cally Russell.
    2. It is based in Edinburgh.

###

Sentence
He was educated at William Ellis School
    before being accepted into
    University College London to study
    botany and zoology, after graduating
     he went to the College of the
    Pharmaceutical Society and studied
    pharmacy, graduating in 1935.

Propositions
    1. He was educated at William Ellis
        School.
    2. This was before being accepted
        into University College London.
    3. This was to study botany and
        zoology.
    4. After graduating he went to the
        College of the Pharmaceutical
        Society.
    5. He studied pharmacy.
    6. He graduated in 1935.

###

Sentence
Out of 3 other surrounding neighborhoods
    , Mattapan saw a population decrease
     but has the highest proportion of
    Black/African American residents in
    the city, but the number of blacks
    actually dropped over the last
    decade.

Propositions
    1. Out of 3 other surrounding
        neighborhoods.
    2. Mattapan saw a population
        decrease.
    3. It has the highest proportion of
        Black/African American residents
        in the city.
    4. The number of blacks actually
        dropped over the last decade.

###

Sentence
Nerepis is situated on the Nerepis River
     and is located east of the town of
    Grand Bay-Westfield in the Saint
    John, the nearest city, which is
    about twenty-five minutes away.

Propositions
    1. Nerepis is situated on the
        Nerepis River.
    2. It is located east of the town of
         Grand Bay-Westfield.
    3. Grand Bay-Westfield is in the
        Saint John.
    4. Saint John is the nearest city.
    5. It is about twenty-five minutes
        from Nerepis.

###

Sentence
In 1961, when Muskee was 20 years old,
    his mother died, and a year later
    his grandmother died.

Propositions
    1. In 1961, when Muskee was 20 years
         old.
    2. His mother died.
    3. A year later, his grandmother
        died.

###

Sentence
{}
```

Paraphrase

```
We would like to propose a list of
    paraphrases of sentences. For each
    sentence, you will provide four
    paraphrases that have the same
    meaning as the original sentence and
     mostly use the same words as well.
Ensure that each of the four paraphrases
     is a correct sentence and does not
    change the meaning of the original
    sentence.
Here are some examples.

###

Sentence
The Boolean satisfiability problem is a
    well-researched problem with many
    exemplar solvers available; it is
    very fast, as package solving
    complexity is very low compared to
    other areas where SAT solvers are
    used.

Propositions
    1. The Boolean satisfiability
```

```
        problem is a widely studied
        topic, with numerous exemplar
        solvers available; it is
        efficient, as solving package
        complexity is significantly
        lower than in other domains
        using SAT solvers.
    2. Boolean satisfiability, a well-
        researched problem, boasts many
        exemplar solvers, and its speed
        is notable due to the low
        complexity of package solving
        compared to other SAT
        applications.
    3. The problem of Boolean
        satisfiability has been
        extensively researched, leading
        to the development of many
        exemplar solvers; package
        solving in this context is fast,
         given its comparatively low
        complexity in contrast to other
        SAT solver uses.
    4. With numerous exemplar solvers
        available, the Boolean
        satisfiability problem is well-
        researched and demonstrates
        remarkable speed, as the
        complexity of package solving is
         much lower than in other SAT
        solver applications.

###

Sentence
Dore was offered several one-off shows
    in night clubs, and her best album
    was rereleased in 2001.

Propositions
    1. Dore's best album was rereleased
        in 2001, and she was offered
        several one-off shows in night
        clubs.
    2. In 2001, Dore's best album was
        rereleased, and she received
        offers for several one-off
        performances in night clubs.
    3. Several one-off shows in night
        clubs were offered to Dore, and
        her best album saw a rerelease
        in 2001.
    4. Dore was given opportunities for
        one-off performances in night
        clubs, and her best album was
        rereleased during 2001.

###

Sentence
Jim briefly transfers to the Stamford
    branch after Pam confirmed her
    commitment to Roy, before corporate
    is forced to merge the Stamford
    branch and staff into the Scranton
    branch.

Propositions
    1. After Pam confirmed her
        commitment to Roy, Jim briefly
        transfers to the Stamford branch
        , only for corporate to merge
        Stamford staff into the Scranton
         branch.
    2. Jim transfers briefly to the
        Stamford branch after Pam
        confirms her commitment to Roy,
        but corporate later merges the
        Stamford staff into the Scranton
         branch.
    3. Pam's confirmation of her
        commitment to Roy leads Jim to
        briefly transfer to the Stamford
         branch, which is later merged
        into the Scranton branch by
        corporate.
    4. Before corporate merges the
        Stamford branch and its staff
        into the Scranton branch, Jim
        briefly transfers there after
        Pam confirms her commitment to
        Roy.

###

Sentence
But Jack could not get back to his own
    time, because one of the drug vials
    had broke, and there was only enough
     left in one of the vials to stop
    Whistler.

Propositions
    1. Jack could not return to his own
        time because one of the drug
        vials had broken, leaving only
        enough in one vial to stop
        Whistler.
    2. Since one of the drug vials had
        broken, Jack was unable to get
        back to his own time, with just
        enough remaining in a single
        vial to stop Whistler.
    3. Because one of the vials of the
        drug had broken, Jack could not
        make it back to his own time, as
         only one vial had enough left
        to stop Whistler.
    4. One of the drug vials had broken,
         leaving Jack unable to return
        to his own time, with only
        enough left in one vial to stop
        Whistler.

###

Sentence
{}
```

C.3.3 Merge prompt

Vanilla

```
Given the following sentence-translation
     pairs written by a professional
    translator:

<Demonstrations>
1. English sentence
The mice used to be diabetic.
```

```
Amharic translation
<>

2. English sentence
They now have 4-month-old mice.
Amharic translation
<>

3. English sentence
The mice are non-diabetic.
Amharic translation
<>
</Demonstrations>

Please write a high-quality Amharic
    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 make sure to consider the above
    information and provide only the
    translation, nothing more
```

C.4 About the decomposition step

For structural decomposition, we use the dependency tree to recursively split the sentence. First, we identify the root of the sentence and divide it into two parts: the left part (including the root) and the right part (excluding the root). Both parts are added to a stack, ensuring that the longer segment (in terms of word count) is processed first. This process continues until all subparts contain no more than four words.

C.5 About the existing strategies

MAPS MAPS (He et al., 2024) is an ensembling strategy where an LLM generates three different translations of a given sentence by analyzing 3 aspects. The available implementation only supports HRLs (English, Chinese, French, German, Japanese). In order to extend it to more languages, we translated their list of keywords using `NLLB-200-distilled-600M`. We found that their trigger sentences came from FLORES-200, so we used their equivalents in other FLORES languages. We also selected demonstrations from FLORES and ensured they were related to the trigger sentences.

SBYS SBYS (Briakou et al., 2024) uses a multi-turn conversation in order to drive an LLM to output a better translation. The authors did not provide an open-source implementation of their method so we had to build it from scratch with the help of the prompts provided in the paper.

TEaR We re-implemented TEaR (Feng et al., 2024) using the prompts provided in their paper to make it compatible with our benchmarks. Since CompTra uses BM25 as its retriever, we used it for TEaR as well. We used Few-shot Translate, Few-shot Estimate and $\mathcal{T}_{refine} - \beta$ as the Refine prompt.

Self-Refine Following Chen et al. (2024), we used their *Refine* prompt which simply ask an LLM to refine a translation given the source sentence. They reported that it usually works best, reaching its peak performance after one iteration of refinement.

CoT For zero-shot CoT, we adopted the traditional *Let's think step by step* (Kojima et al., 2022) trigger sentence with the two-step procedure: Reasoning extraction and answer extraction. In few-shot CoT, we did not provide examples of step-by-step reasoning in the demonstrations, instead we just appended *Let's think step by step* at the end of the prompt (after the few-shot demonstrations) and did the two-step procedure as in zero-shot CoT.

D Qualitative Comparison between the outputs of models

Source sentence	His work is of such recognized quality and detail that he is one of the very few "household names" among philatelists. Some specialize in collecting his work alone.
Zero-shot	O le galuega a ia o le aofaiga ma le faamalieina o le faamaonia, o lea e faaigoa ai o ia i totonu o le laupesi o tagata faatau tupe. E toaitiiti tagata e faaaluina o latou galuega i le faatauina o ana galuega e tasi.
Zero-shot + CoT	O le galuega a ia e fa'amaonia lelei ma le fa'apalepale o lea e tasi o le vae to'atele o tagata e fa'amaonia i totonu o tagata fa'afofoga pepa. E iai ni tagata e fa'afofoga i le galuega a ia e tasi.
Zero-shot + Refine	O le galuega a ia o le aofaiga ma le faamalieina o le faamaonia, o lea e faaigoa ai o ia i totonu o le laupesi o tagata faatau tupe. E toaitiiti tagata e faaaluina o latou galuega i le faatauina o ana galuega e tasi.
SBYS	O le galuega a ia o le aofaiga ma le faamalieina o le tulaga e faamaonia, o le mea lea e faaigoa ai o ia i totonu o le au faatau tupe, e lei tele ni tagata e faatau tupe e faaalu i le galuega a ia.
MAPS	O lana galuega e faamaonia lelei ma le faamalieina o le aiga o le igoa o le aiga i totonu o tagata faatau pese. E toaitiiti tagata e faapipii i le faatauina o lana galuega e tasi.
TEaR	O lana galuega e faamaonia lelei ma le faamalieina, o le mea lea e faaalu ai o ia i totonu o le vae toaitiiti o "igoa o le aiga" i totonu o tagata faatau tupe. E toaitiiti e faapipii i le faatauina o lana galuega e aunoa.
5-shot BM25	O lana galuega e faamaonia lelei ma le faamalieina o le aofai, o le mea lea e faaigoa ai o ia o se tasi o le vae toaitiiti o "igoa o fale" i totonu o tagata faatau tupe. E toaitiiti e faaoga i le faatauina o lana galuega e tasi.
5-shot BM25 + CoT	Lana galuega e fa'amaonia ma le fa'afine, o le mea lea e fa'apea o ia o se tasi o le vae to'aitiiti o igoa o le aiga i totonu o tagata fa'afoe. E iai nisi e fa'aaluina i le fa'atauina o lana galuega e tasi.
5-shot BM25 + Refine	O lana galuega e faamaonia lelei ma le faamalieina o le aofai, o le mea lea e faaigoa ai o ia o se tasi o tagata e sili ona lauiloa i totonu o tagata faatau tupe. E toaitiiti e faaoga i le faatauina o lana galuega e tasi.
CompTra	O lana galuega e fa'amaonia lelei ma le fa'afouina, o se tasi o ni vaaiga lauiloa i totonu o tagata fa'aalu i tupe. E i ai nisi e sili ona fa'aauau i le fa'asoa o lana galuega e tasi.
Reference	O lana galuega sa lauiloa i le tulaga lelei ma sa lauiloa lona igoa i le lisi o e faia faailoga o tusi (stamps). E i ai tagata faapitoa i le aoina mai ana galuega.

Table 28: `LLaMA-3.1-70B-It`'s generations in Samoan.