

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

Armel Zebaze¹ Benoît Sagot¹ Rachel Bawden¹

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>.

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. Multi-

¹ Inria, Paris, France. Correspondence to: Armel Zebaze <armel.zebaze-dongmo@inria.fr>.

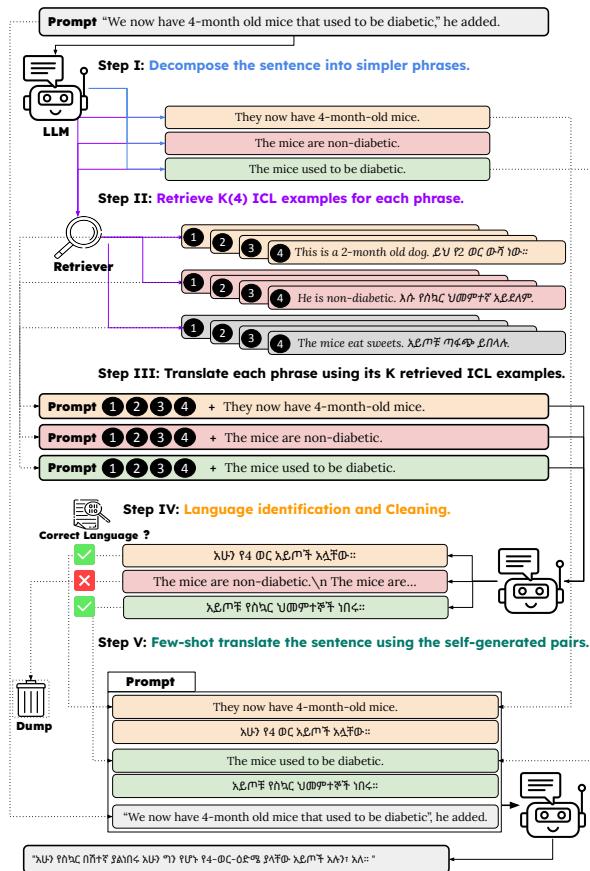


Figure 1. An overview of **Compositional Translation (CompTra)**. Given a sentence, our method prompts the LLM to decompose it into several phrases that use its words. For each phrase, we retrieve relevant in-context demonstrations (here four) through similarity search and use them to translate it in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

ple 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 machine translation (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

sentence to translate (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 property of compositionality of MT (Turcato & 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 & 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 treated as a complex step by step reasoning task 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, yielding 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 size 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 few-shot, with demonstrations retrieved via similarity-search in a selection pool (tightening the constraint on the access to outside knowledge). Finally, the LLM derives the final translation by drawing insights from the above 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. Moreover, 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 consistently 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 & 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); 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 candidate pool offers no significant benefit. In contrast, Zebaze et al. (2024a) highlighted that while similarity-based selection does not help HRLs, it significantly improves performance for LRLs.

Prompting and Compositionality After Brown et al. (2020) demonstrated the ability of LLMs to use ICL on diverse tasks, the research community investigated the development of better performing strategies for problem-solving at inference time. Wei et al. (2022) introduced chain-of-thought (CoT) prompting, which helps 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 the 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 teach-

ing 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 the MT task.

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 translate 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 and thus faster. 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 & 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 (Aycock & Bawden, 2024), 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 & 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; ImaniGooghar et al., 2023; Singh et al., 2024; Futerla et al., 2024), as well as efforts in multilingual and multitask 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 the decomposition is to create **simpler** phrases that share words with the source sentence. We achieve this with the help of 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 filterout 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

4.1. Experimental setup

Datasets. We work on MT from English (eng) to LRLs.

- **FLORES 200** (Goyal et al., 2022; Costa-jussà et al., 2022). This dataset consists of translations from web articles into 204 languages. These sentences are divided into two splits: dev and devtest. We use the FLORES 200 dev set (997 examples) as the selection pool and the FLORES 200 devtest set (1012 examples) for the 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 covering 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 Metrics. We mainly evaluate using MetricX-23 (Juraska et al., 2023) and XCOMET (Guerreiro et al., 2024), which are the highest ranked non-ensemble metrics according to the latest WMT shared task on MT metrics (Freitag et al., 2024). Precisely we use their reference-based versions XCOMET-XXL (which supports the same

¹command-r-08-2024, command-r-plus-08-2024

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 assigns a score ranging from 0 to 25, with higher scores indicating more errors in the translation. XCOMET produces a score between 0 and 1, which we rescale to a range of 0 to 100, where higher scores represent better translation quality. We also consider n -gram matching metrics via sacreBLEU (Post, 2018), namely chrF++² (Popović, 2015; Popović, 2017) and BLEU³ (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 a candidate 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 5 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 (Lü, 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

4.2. Results

4.2.1. MAIN RESULTS: FLORES 200

We evaluate CompTra on 10 English-to-X translation directions from FLORES 200 (Table 1).⁵ CompTra consistently outperforms similarity-based few-shot MT across all directions and for all the LLMs we evaluated. On average, CompTra outperforms few-shot BM25 by 0.4 MetricX with LLaMA 3.1 70B It and Gemma 2 27B It, and by 1.5 MetricX with Command-R+. For XCOMET, the gains are 1.0 with Gemma 2 27B It and 1.8 with Command-R+.

4.2.2. RESULTS ON NTREX 128 AND TICO-19

We report the results obtained by CompTra on NTREX 128 and TICO-19 in Tables 2⁶ and 3 respectively. Similar to our observations on FLORES-200, CompTra consistently outperforms similarity-based few-shot MT across all directions and for all the LLMs.

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

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

⁴See Appendix C.1.

⁵See Appendix B.1 for BLEU and chrF++ scores.

⁶See Appendix B.2 for BLEU and chrF++ scores.

Table 1. Full quantitative results for 10 English → X translation directions from FLORES 200 (Goyal et al., 2022; Costa-jussà et al., 2022) (XCOMET and MetricX scores). Best results (including any results that are not statistically worse) are highlighted in bold.

Methods	Amharic		Burmese		Fijian		Khmer		Lao	
	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								
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 2. Full XCOMET and MetricX results for 5 English→X directions from NTREX 128 (Federmann et al., 2022).

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 3. Full XCOMET and MetricX results for 5 English→X translation directions from TICO-19 (Anastasopoulos et al., 2020).

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

4.2.3. COMPARISON TO EXISTING APPROACHES

We conduct a set of additional studies and compare *compositional translation* 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 Table 4. CompTra significantly outperforms all the strategies. 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.

Table 5. Full MetricX results for ten English→X directions with smaller LMs.

	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

5. Analysis

Does CompTra work with weaker LLMs? In all our experiments in Section 4.2, 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 5 (See Appendix B.3 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 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 TICO-19 test set (COVID-19; health domain) and the selection pool is FLORES 200 dev set (news domain) as opposed to the usual TICO-19 validation set (health domain). As expected and reported in Table 9, in-domain scores are better than their out-of-domain counterparts. However, in both scenarios, applying CompTra yields gains over standalone retrieval-based few-shot MT. This suggests that CompTra can be successfully applied in setups where there is 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 per phrase. As shown in Figure 2, 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++ and 2.5 chrF++ 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 in Samoan with LLaMA 3.1 8B It. Small values of k tend to yield smaller gains and we attribute this to the fact that despite small sentences (phrases) being easier to translate for LLMs,

Table 4. 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	
	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 6. Impact of switching few-shot MT with NLLB in CompTra (MetricX scores).

	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

Table 7. Ablation study on the impact of the retriever on 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 8. Ablation study on the impact of the decomposition on CompTra (MetricX scores).

	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

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 by a direction translation obtained using a supervised translation model, here NLLB-200-distilled-600M (Costa-

Table 9. In-domain and out-of-domain evaluation (MetricX).

	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
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

jussà et al., 2022). The LLM therefore 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 6. **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 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.

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 7 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 Min-WikiSplit 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**

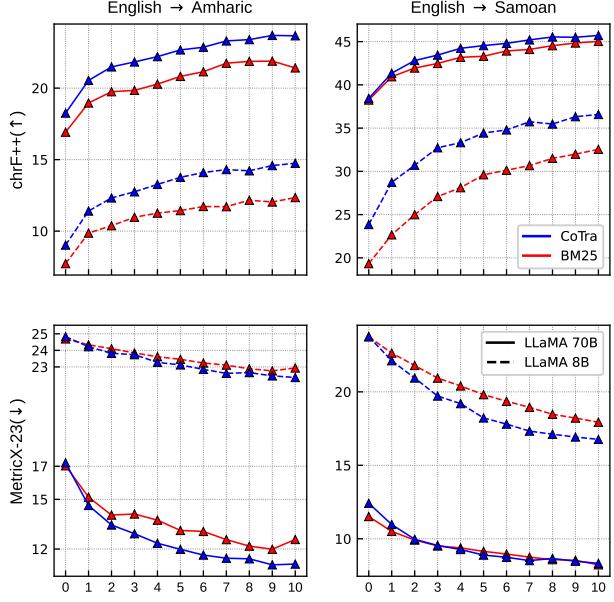


Figure 2. Impact of the number of in-context examples per phrase.

divides the sentence into phrases by heuristically analyzing its dependency tree.⁷ In Table 8, 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 (4.912) than CompTra’s native form (3.166) 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.

⁷See appendix C.4.

Acknowledgements

This work was partly funded by the last two authors' chairs in the PRAIRIE-PSAI institute funded by the French national agency ANR as part of the “France 2030” strategy under the reference ANR-23-IACL-0008. The authors are grateful to the OPAL infrastructure from Université Côte d’Azur for providing resources and support. This work was 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.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Translation with Large Language Models. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

- Abadji, J., Ortiz Suarez, P., Romary, L., and Sagot, B. Towards a cleaner document-oriented multilingual crawled corpus. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Mægaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S. (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4344–4355, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.lrec-1.463/>.
- Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., and Ghazvininejad, M. In-context examples selection for machine translation. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8857–8873, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.564. URL <https://aclanthology.org/2023.findings-acl.564>.
- Anastasopoulos, A., Cattelan, A., Dou, Z.-Y., Federico, M., Federmann, C., Genzel, D., Guzmán, F., Hu, J., Hughes, M., Koehn, P., Lazar, R., Lewis, W., Neubig, G., Niu, M., Öktem, A., Paquin, E., Tang, G., and Tur, S. TICO-19: the translation initiative for COvid-19. In Verspoor, K., Cohen, K. B., Conway, M., de Bruijn, B., Dredze, M., Mihalcea, R., and Wallace, B. (eds.), *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcovid19-2.5. URL <https://aclanthology.org/2020.nlpcovid19-2.5>.
- Aycock, S. and Bawden, R. Topic-guided example selection for domain adaptation in LLM-based machine translation. In Falk, N., Papi, S., and Zhang, M. (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 175–195, St. Julian’s, Malta, March 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.eacl-srw.13/>.
- Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. Findings of the 2019 conference on machine translation (WMT19). In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Turchi, M., and Verspoor, K. (eds.), *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 1–61, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5301. URL <https://aclanthology.org/W19-5301>.
- Bavarian, M., Jun, H., Tezak, N., Schulman, J., McLeavey, C., Tworek, J., and Chen, M. Efficient training of language models to fill in the middle. *arXiv preprint arXiv:2207.14255*, 2022.
- Bawden, R. and Yvon, F. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In Nurminen, M., Brenner, J., Koponen, M., Latomaa, S., Mikhailov, M., Schierl, F., Ranasinghe, T., Vanmassenhove, E., Vidal, S. A., Aranberri, N., Nunziatini, M., Escartín, C. P., Forcada, M., Popovic, M., Scarton, C., and Moniz, H. (eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 157–170, Tampere, Finland, June 2023. European Association for Machine Translation. URL <https://aclanthology.org/2023.eamt-1.16>.
- Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawska, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., and Hoefer, T. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690, Mar. 2024. doi: 10.1609/aaai.v38i16.29720. URL <https://ojs.aaai.org/index.php/AAAI/article/view/29720>.
- BigScience Workshop, Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., Ruwase,

- O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P. O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A. F., Alfassy, A., Rogers, A., Nitzav, A. K., Xu, C., Mou, C., Emezue, C., Klamm, C., Leong, C., van Strien, D., Adelani, D. I., Radev, D., Ponferrada, E. G., Levkovizh, E., Kim, E., Natan, E. B., Toni, F. D., Dupont, G., Kruszewski, G., Pistilli, G., Elsahar, H., Benyamina, H., Tran, H., Yu, I., Abdulmumin, I., Johnson, I., Gonzalez-Dios, I., de la Rosa, J., Chim, J., Dodge, J., Zhu, J., Chang, J., Frohberg, J., Tobing, J., Bhattacharjee, J., Almubarak, K., Chen, K., Lo, K., Werra, L. V., Weber, L., Phan, L., allal, L. B., Tanguy, L., Dey, M., Muñoz, M. R., Masoud, M., Grandury, M., Šaško, M., Huang, M., Coavoux, M., Singh, M., Jiang, M. T.-J., Vu, M. C., Jauhar, M. A., Ghaleb, M., Subramani, N., Kassner, N., Khamis, N., Nguyen, O., Espejel, O., de Gibert, O., Villegas, P., Henderson, P., Colombo, P., Amuok, P., Lhoest, Q., Harliman, R., Bommasani, R., López, R. L., Ribeiro, R., Osei, S., Pyysalo, S., Nagel, S., Bose, S., Muhammad, S. H., Sharma, S., Longpre, S., Nikpoor, S., Silberberg, S., Pai, S., Zink, S., Torrent, T. T., Schick, T., Thrush, T., Danchev, V., Nikoulin, V., Laippala, V., Lepercq, V., Prabhu, V., Alyafeai, Z., Talat, Z., Raja, A., Heinzerling, B., Si, C., Taşar, D. E., Salesky, E., Mielke, S. J., Lee, W. Y., Sharma, A., Santilli, A., Chaffin, A., Stiegler, A., Datta, D., Szczechla, E., Chhablani, G., Wang, H., Pandey, H., Strobel, H., Fries, J. A., Rozen, J., Gao, L., Sutawika, L., Bari, M. S., Al-shaibani, M. S., Manica, M., Nayak, N., Teehan, R., Albanie, S., Shen, S., Ben-David, S., Bach, S. H., Kim, T., Bers, T., Fevry, T., Neeraj, T., Thakker, U., Raunak, V., Tang, X., Yong, Z.-X., Sun, Z., Brody, S., Uri, Y., Tojarieh, H., Roberts, A., Chung, H. W., Tae, J., Phang, J., Press, O., Li, C., Narayanan, D., Bourfoune, H., Casper, J., Rasley, J., Ryabinin, M., Mishra, M., Zhang, M., Shoeybi, M., Peyrounette, M., Patry, N., Tazi, N., Sansevieri, O., von Platen, P., Cornette, P., Lavallée, P. F., Lacroix, R., Rajbhandari, S., Gandhi, S., Smith, S., Requena, S., Patil, S., Dettmers, T., Baruwa, A., Singh, A., Cheveleva, A., Ligozat, A.-L., Subramonian, A., Névéol, A., Lovering, C., Garrette, D., Tunuguntla, D., Reiter, E., Taktasheva, E., Voloshina, E., Bogdanov, E., Winata, G. I., Schoelkopf, H., Kalo, J.-C., Novikova, J., Forde, J. Z., Clive, J., Kasai, J., Kawamura, K., Hazan, L., Carpuat, M., Clinciu, M., Kim, N., Cheng, N., Serikov, O., Antverg, O., van der Wal, O., Zhang, R., Zhang, R., Gehrmann, S., Mirkin, S., Pais, S., Shavrina, T., Scialom, T., Yun, T., Limisiewicz, T., Rieser, V., Protasov, V., Mikhailov, V., Prusachatkun, Y., Belinkov, Y., Bamberger, Z., Kasner, Z., Rueda, A., Pestana, A., Feizpour, A., Khan, A., Faranak, A., Santos, A., Hevia, A., Unldreaj, A., Aghagol, A., Abdollahi, A., Tammour, A., HajiHosseini, A., Behroozi, B., Ajibade, B., Saxena, B., Ferrandis, C. M., McDuff, D., Contractor, D., Lansky, D., David, D., Kiela, D., Nguyen, D. A., Tan, E., Baylor, E., Ozoani, E., Mirza, F., Ononiwu, F., Rezanejad, H., Jones, H., Bhattacharya, I., Solaiman, I., Sedenko, I., Nejadgholi, I., Passmore, J., Seltzer, J., Sanz, J. B., Dutra, L., Samagaio, M., Elbadri, M., Mieskes, M., Gerchick, M., Akinalolu, M., McKenna, M., Qiu, M., Ghauri, M., Burynok, M., Abrar, N., Rajani, N., Elkott, N., Fahmy, N., Samuel, O., An, R., Kromann, R., Hao, R., Alizadeh, S., Shubber, S., Wang, S., Roy, S., Viguer, S., Le, T., Oyebade, T., Le, T., Yang, Y., Nguyen, Z., Kashyap, A. R., Palasciano, A., Callahan, A., Shukla, A., Miranda-Escalada, A., Singh, A., Beilharz, B., Wang, B., Brito, C., Zhou, C., Jain, C., Xu, C., Fourrier, C., Periñán, D. L., Molano, D., Yu, D., Manjavacas, E., Barth, F., Fuhrmann, F., Altay, G., Bayrak, G., Burns, G., Vrabec, H. U., Bello, I., Dash, I., Kang, J., Giorgi, J., Golde, J., Posada, J. D., Sivaraman, K. R., Bulchandani, L., Liu, L., Shinzato, L., de Bykhovetz, M. H., Takeuchi, M., Pàmies, M., Castillo, M. A., Nezhurina, M., Sänger, M., Samwald, M., Cullan, M., Weinberg, M., Wolf, M. D., Mihaljcic, M., Liu, M., Freidank, M., Kang, M., Seelam, N., Dahlberg, N., Broad, N. M., Muellner, N., Fung, P., Haller, P., Chandrasekhar, R., Eisenberg, R., Martin, R., Canalli, R., Su, R., Su, R., Cahyawijaya, S., Garda, S., Deshmukh, S. S., Mishra, S., Kiblawi, S., Ott, S., Sang-aroonisiri, S., Kumar, S., Schweter, S., Bharati, S., Laud, T., Gigant, T., Kainuma, T., Kusa, W., Labrak, Y., Bajaj, Y. S., Venkatraman, Y., Xu, Y., Xu, Y., Xu, Y., Tan, Z., Xie, Z., Ye, Z., Bras, M., Belkada, Y., and Wolf, T. Bloom: A 176b-parameter open-access multilingual language model, 2023.
- Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi: 10.1162/tacl_a_00051. URL <https://aclanthology.org/Q17-1010>.
- Bouthors, M., Crego, J., and Yvon, F. Retrieving examples from memory for retrieval augmented neural machine translation: A systematic comparison. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3022–3039, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.190. URL <https://aclanthology.org/2024.findings-naacl.190>.
- Briakou, E., Luo, J., Cherry, C., and Freitag, M. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In *Proceedings of the Ninth Conference on Machine Translation*, pp. 1301–1317, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.wmt-1.123>.

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.
- Cahyawijaya, S., Lovenia, H., and Fung, P. LLMs are few-shot in-context low-resource language learners. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 405–433, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.24. URL <https://aclanthology.org/2024.naacl-long.24>.
- Chen, M., Duquenne, P.-A., Andrews, P., Kao, J., Mourachko, A., Schwenk, H., and Costa-jussà, M. R. BLASER: A text-free speech-to-speech translation evaluation metric. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9064–9079, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.504. URL <https://aclanthology.org/2023.acl-long.504>.
- Chen, P., Guo, Z., Haddow, B., and Heafield, K. Iterative translation refinement with large language models, 2024. URL <https://arxiv.org/abs/2306.03856>.
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pelлат, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL <http://jmlr.org/papers/v24/22-1144.html>.
- Cohere. Command r+, 2024. <https://docs.cohere.com/v2/docs/command-r-plus>.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL <https://aclanthology.org/2020.acl-main.747>.
- Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A. Y., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H., and Wang, J. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672, 2022. doi: 10.48550/ARXIV.2207.04672. URL <https://doi.org/10.48550/arXiv.2207.04672>.
- Dou, L., Liu, Q., Zeng, G., Guo, J., Zhou, J., Lu, W., and Lin, M. Sailor: Open language models for south-east asia, 2024. URL <https://arxiv.org/abs/2404.03608>.
- Doumbouya, M., Diané, B. M., Cissé, S. F., Diané, D., Sow, A., Doumbouya, S. M., Bangoura, D., Bayo, F. M., Conde, I. S., Diané, K. M., Piech, C., and Manning, C. Machine translation for nko: Tools, corpora, and baseline results. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C. (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 312–343, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.34. URL <https://aclanthology.org/2023.wmt-1.34>.
- Dua, D., Gupta, S., Singh, S., and Gardner, M. Successive prompting for decomposing complex questions. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1251–1265, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.81. URL <https://aclanthology.org/2022.emnlp-main.81>.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A.,

- Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Al-lonsius, D., Song, D., Pintz, D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Al-wala, K. V., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Rantala-Yearly, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Tan, X. E., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Grattafiori, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenber, A., Vaughan, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Franco, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Wyatt, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Ozgenel, F., Caggioni, F., Guzmán, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Thattai, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Molybog, I., Tufanov, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Prasad, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Huang, K., Chawla, K., Lakhotia, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Tsim-poukelli, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Laptev, N. P., Dong, N., Zhang, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Li, R., Hogan, R., Battey, R., Wang, R., Maheswari, R., Howes, R., Rinott, R., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Kohler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wang, X., Wu, X., Wang, X., Xia, X., Wu, X., Gao, X., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang,

- Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Hao, Y., Qian, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., and Zhao, Z. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Duquenne, P.-A., Schwenk, H., and Sagot, B. Sonar: Sentence-level multimodal and language-agnostic representations, 2023. URL <https://arxiv.org/abs/2308.11466>.
- Enis, M. and Hopkins, M. From llm to nmt: Advancing low-resource machine translation with claude, 2024. URL <https://arxiv.org/abs/2404.13813>.
- Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., Chaudhary, V., Goyal, N., Birch, T., Liptchinsky, V., Edunov, S., Auli, M., and Joulin, A. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48, 2021. URL <http://jmlr.org/papers/v22/20-1307.html>.
- Federmann, C., Kocmi, T., and Xin, Y. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pp. 21–24, Online, nov 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.sumeval-1.4>.
- Feng, Z., Zhang, Y., Li, H., Wu, B., Liao, J., Liu, W., Lang, J., Feng, Y., Wu, J., and Liu, Z. Tear: Improving llm-based machine translation with systematic self-refinement, 2024. URL <https://arxiv.org/abs/2402.16379>.
- Freitag, M., Grangier, D., and Caswell, I. BLEU might be guilty but references are not innocent. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 61–71, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.5. URL <https://aclanthology.org/2020.emnlp-main.5>.
- Freitag, M., Mathur, N., Deutsch, D., Lo, C.-k., Avramidis, E., Rei, R., Thompson, B., Blain, F., Kocmi, T., Wang, J., Adelani, D. I., Buchicchio, M., Zerva, C., and Lavie, A. Are llms breaking mt metrics? results of the wmt24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pp. 47–81, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.wmt-1.2>.
- Fu, Y., Peng, H., Sabharwal, A., Clark, P., and Khot, T. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=yflicZHC-19>.
- Futeral, M., Zebaze, A., Suarez, P. O., Abadji, J., Lacroix, R., Schmid, C., Bawden, R., and Sagot, B. moscar: A large-scale multilingual and multimodal document-level corpus, 2024. URL <https://arxiv.org/abs/2406.08707>.
- Gemini Team, Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., Mariooryad, S., Ding, Y., Geng, X., Alcober, F., Frostig, R., Omernick, M., Walker, L., Paduraru, C., Sorokin, C., Tacchetti, A., Gaffney, C., Daruki, S., Sercinoglu, O., Gleicher, Z., Love, J., Voigtlaender, P., Jain, R., Surita, G., Mohamed, K., Blevins, R., Ahn, J., Zhu, T., Kawintiranon, K., Firat, O., Gu, Y., Zhang, Y., Rahtz, M., Faruqui, M., Clay, N., Gilmer, J., Co-Reyes, J., Penchev, I., Zhu, R., Morioka, N., Hui, K., Haridasan, K., Campos, V., Mahdieh, M., Guo, M., Hassan, S., Kilgour, K., Vezer, A., Cheng, H.-T., de Liedekerke, R., Goyal, S., Barham, P., Strouse, D., Noury, S., Adler, J., Sundararajan, M., Vikram, S., Lepikhin, D., Paganini, M., Garcia, X., Yang, F., Valter, D., Trebacz, M., Vodrahalli, K., Asawaroengchai, C., Ring, R., Kalb, N., Soares, L. B., Brahma, S., Steiner, D., Yu, T., Mentzer, F., He, A., Gonzalez, L., Xu, B., Kaufman, R. L., Shafey, L. E., Oh, J., Hennigan, T., van den Driessche, G., Odoom, S., Lucic, M., Roelofs, B., Lall, S., Marathe, A., Chan, B., Ontanon, S., He, L., Teplyashin, D., Lai, J., Crone, P., Damoc, B., Ho, L., Riedel, S., Lenc, K., Yeh, C.-K., Chowdhery, A., Xu, Y., Kazemi, M., Amid, E., Petrushkina, A., Swersky, K., Khodaei, A., Chen, G., Larkin, C., Pinto, M., Yan, G., Badia, A. P., Patil, P., Hansen, S., Orr, D., Arnold, S. M. R., Grimstad, J., Dai, A., Douglas, S., Sinha, R., Yadav, V., Chen, X., Gribovskaya, E., Austin, J., Zhao, J., Patel, K., Komarek, P., Austin, S., Borgeaud, S., Friso, L., Goyal, A., Caine, B., Cao, K., Chung, D.-W., Lamm, M., Barth-Maron, G., Kagohara, T., Olszewska, K., Chen, M., Shivakumar, K., Agarwal, R., Godhia, H., Rajwar, R., Snaider, J., Dotiwala, X., Liu, Y., Barua, A., Ungureanu, V., Zhang, Y., Batsaikhan, B.-O., Wirth, M., Qin, J., Danihelka, I., Doshi, T., Chadwick, M., Chen, J., Jain, S., Le, Q., Kar, A., Gurumurthy, M., Li, C., Sang, R., Liu, F., Lamprou, L., Munoz, R., Lintz, N., Mehta, H., Howard, H., Reynolds, M., Aroyo, L., Wang, Q., Blanco, L., Cassirer, A., Griffith, J., Das, D., Lee, S., Sygnowski, J., Fisher, Z., Besley, J., Powell, R., Ahmed, Z., Paulus, D., Reitter, D., Borsos, Z., Joshi, R., Pope, A., Hand, S., Selo, V., Jain, V., Sethi, N., Goel, M., Makino, T., May, R., Yang, Z., Schalkwyk, J., Butterfield, C., Hauth, A., Goldin, A., Hawkins, W., Senter, E., Brin, S., Woodman, O., Ritter, M., Noland, E., Giang, M., Bolina, V., Lee, L., Blyth, T., Mackinnon, I., Reid, M., Sarvana, O., Silver, D., Chen, A., Wang, L., Maggiore, L., Chang, O., At-

taluri, N., Thornton, G., Chiu, C.-C., Bunyan, O., Levine, N., Chung, T., Eltyshev, E., Si, X., Lillicrap, T., Brady, D., Aggarwal, V., Wu, B., Xu, Y., McIlroy, R., Badola, K., Sandhu, P., Moreira, E., Stokowiec, W., Hemsley, R., Li, D., Tudor, A., Shyam, P., Rahimtoroghi, E., Haykal, S., Sprechmann, P., Zhou, X., Mincu, D., Li, Y., Addanki, R., Krishna, K., Wu, X., Frechette, A., Eyal, M., Dafoe, A., Lacey, D., Whang, J., Avrahami, T., Zhang, Y., Taropa, E., Lin, H., Toyama, D., Rutherford, E., Sano, M., Choe, H., Tomala, A., Safranek-Shrader, C., Kassner, N., Pajarskas, M., Harvey, M., Sechrist, S., Fortunato, M., Lyu, C., Elsayed, G., Kuang, C., Lottes, J., Chu, E., Jia, C., Chen, C.-W., Humphreys, P., Baumli, K., Tao, C., Samuel, R., dos Santos, C. N., Andreassen, A., Rakićević, N., Grewe, D., Kumar, A., Winkler, S., Caton, J., Brock, A., Dalmia, S., Sheahan, H., Barr, I., Miao, Y., Natsev, P., Devlin, J., Behbahani, F., Prost, F., Sun, Y., Myaskovsky, A., Pillai, T. S., Hurt, D., Lazaridou, A., Xiong, X., Zheng, C., Pardo, F., Li, X., Horgan, D., Stanton, J., Ambar, M., Xia, F., Lince, A., Wang, M., Mustafa, B., Webson, A., Lee, H., Anil, R., Wicke, M., Dozat, T., Sinha, A., Piqueras, E., Dabir, E., Upadhyay, S., Boral, A., Hendricks, L. A., Fry, C., Djolonga, J., Su, Y., Walker, J., Labanowski, J., Huang, R., Misra, V., Chen, J., Skerry-Ryan, R., Singh, A., Rijhwani, S., Yu, D., Castro-Ros, A., Changpinyo, B., Datta, R., Bagri, S., Hrafnelsson, A. M., Maggioni, M., Zheng, D., Sulsky, Y., Hou, S., Paine, T. L., Yang, A., Riesa, J., Rogozinska, D., Marcus, D., Badawy, D. E., Zhang, Q., Wang, L., Miller, H., Greer, J., Sjos, L. L., Nova, A., Zen, H., Chaabouni, R., Rosca, M., Jiang, J., Chen, C., Liu, R., Sainath, T., Krikun, M., Polozov, A., Lespiau, J.-B., Newlan, J., Cankara, Z., Kwak, S., Xu, Y., Chen, P., Coenen, A., Meyer, C., Tsihlas, K., Ma, A., Gottweis, J., Xing, J., Gu, C., Miao, J., Frank, C., Cankara, Z., Ganapathy, S., Dasgupta, I., Hughes-Fitt, S., Chen, H., Reid, D., Rong, K., Fan, H., van Amersfoort, J., Zhuang, V., Cohen, A., Gu, S. S., Mohananey, A., Ilic, A., Tobin, T., Wieting, J., Bortsova, A., Thacker, P., Wang, E., Caveness, E., Chiu, J., Sezener, E., Kaskasoli, A., Baker, S., Millican, K., Elhawaty, M., Aisopos, K., Lebsack, C., Byrd, N., Dai, H., Jia, W., Wiethoff, M., Davoodi, E., Weston, A., Yagati, L., Ahuja, A., Gao, I., Pundak, G., Zhang, S., Azzam, M., Sim, K. C., Caelles, S., Keeling, J., Sharma, A., Swing, A., Li, Y., Liu, C., Bostock, C. G., Bansal, Y., Nado, Z., Anand, A., Lipschultz, J., Karmarkar, A., Proleev, L., Ittycheriah, A., Yeganeh, S. H., Polovets, G., Faust, A., Sun, J., Rrustemi, A., Li, P., Shivanna, R., Liu, J., Welty, C., Lebron, F., Baddepudi, A., Krause, S., Parisotto, E., Soricut, R., Xu, Z., Bloxwich, D., Johnson, M., Neyshabur, B., Mao-Jones, J., Wang, R., Ramasesh, V., Abbas, Z., Guez, A., Segal, C., Nguyen, D. D., Svensson, J., Hou, L., York, S., Milan, K., Bridgers, S., Gworek, W., Tagliasacchi, M., Lee-Thorp, J., Chang, M., Guseynov, A., Hartman, A. J., Kwong, M., Zhao, R., Kashem, S., Cole, E., Miech, A., Tanburn, R., Phuong, M., Pavetic, F., Cevey, S., Comanescu, R., Ives, R., Yang, S., Du, C., Li, B., Zhang, Z., Inuma, M., Hu, C. H., Roy, A., Bijwadia, S., Zhu, Z., Martins, D., Saputro, R., Gergely, A., Zheng, S., Jia, D., Antonoglou, I., Sadovsky, A., Gu, S., Bi, Y., Andreev, A., Samangooei, S., Khan, M., Kociský, T., Filos, A., Kumar, C., Bishop, C., Yu, A., Hodkinson, S., Mittal, S., Shah, P., Moufarek, A., Cheng, Y., Bloniarz, A., Lee, J., Pejman, P., Michel, P., Spencer, S., Feinberg, V., Xiong, X., Savinov, N., Smith, C., Shakeri, S., Tran, D., Chesus, M., Bohnet, B., Tucker, G., von Glehn, T., Muir, C., Mao, Y., Kazawa, H., Slone, A., Soparkar, K., Shrivastava, D., Cobon-Kerr, J., Sharman, M., Pavagadhi, J., Araya, C., Misiunas, K., Ghelani, N., Laskin, M., Barker, D., Li, Q., Briukhov, A., Houlsby, N., Glaese, M., Lakshminarayanan, B., Schucher, N., Tang, Y., Collins, E., Lim, H., Feng, F., Recasens, A., Lai, G., Magni, A., Cao, N. D., Siddhant, A., Ashwood, Z., Orbay, J., Dehghani, M., Brennan, J., He, Y., Xu, K., Gao, Y., Saroufim, C., Molloy, J., Wu, X., Arnold, S., Chang, S., Schrittweiser, J., Buchatskaya, E., Radpour, S., Polacek, M., Giordano, S., Bapna, A., Tokumine, S., Hellendoorn, V., Sotiaux, T., Cogan, S., Severyn, A., Saleh, M., Thakoor, S., Shefey, L., Qiao, S., Gaba, M., yiin Chang, S., Swanson, C., Zhang, B., Lee, B., Rubenstein, P. K., Song, G., Kwiatkowski, T., Koop, A., Kannan, A., Kao, D., Schuh, P., Stjerngren, A., Ghiasi, G., Gibson, G., Vilnis, L., Yuan, Y., Ferreira, F. T., Kamath, A., Klimenko, T., Franko, K., Xiao, K., Bhattacharya, I., Patel, M., Wang, R., Morris, A., Strudel, R., Sharma, V., Choy, P., Hashemi, S. H., Landon, J., Finkelstein, M., Jhakra, P., Frye, J., Barnes, M., Mauger, M., Daun, D., Baatarsukh, K., Tung, M., Farhan, W., Michalewski, H., Viola, F., de Chaumont Quiry, F., Lan, C. L., Hudson, T., Wang, Q., Fischer, F., Zheng, I., White, E., Dragan, A., baptiste Alayrac, J., Ni, E., Pritzel, A., Iwanicki, A., Isard, M., Bulanova, A., Zilka, L., Dyer, E., Sachan, D., Srinivasan, S., Muckenhirn, H., Cai, H., Mandhane, A., Tariq, M., Rae, J. W., Wang, G., Ayoub, K., FitzGerald, N., Zhao, Y., Han, W., Alberti, C., Garrette, D., Krishnakumar, K., Gimenez, M., Levskaya, A., Sohn, D., Mata, J., Iturrat, I., Chang, M. B., Xiang, J., Cao, Y., Ranka, N., Brown, G., Hutter, A., Mirrokni, V., Chen, N., Yao, K., Egyed, Z., Galilee, F., Liechty, T., Kallakuri, P., Palmer, E., Ghemawat, S., Liu, J., Tao, D., Thornton, C., Green, T., Jasarevic, M., Lin, S., Cotruta, V., Tan, Y.-X., Fiedel, N., Yu, H., Chi, E., Neitz, A., Heitkaemper, J., Sinha, A., Zhou, D., Sun, Y., Kaed, C., Hulse, B., Mishra, S., Georgaki, M., Kudugunta, S., Farabet, C., Shafran, I., Vlasic, D., Tsitsulin, A., Ananthanarayanan, R., Carin, A., Su, G., Sun, P., V. S., Carvajal, G., Broder, J., Comsa, I., Repina, A., Wong, W., Chen, W. W., Hawkins, P., Filonov, E., Loher, L., Hirnschall, C., Wang, W., Ye, J., Burns, A., Cate, H., Wright, D. G., Piccinini, F., Zhang, L., Lin, C.-C., Gog, I., Kulizhskaya, M.

- Y., Sreevatsa, A., Song, S., Cobo, L. C., Iyer, A., Tekur, C., Garrido, G., Xiao, Z., Kemp, R., Zheng, H. S., Li, H., Agarwal, A., Ngani, C., Goshvadi, K., Santamaria-Fernandez, R., Fica, W., Chen, X., Gorgolewski, C., Sun, S., Garg, R., Ye, X., Eslami, S. M. A., Hua, N., Simon, J., Joshi, P., Kim, Y., Tenney, I., Potluri, S., Thiet, L. N., Yuan, Q., Luisier, F., Chronopoulou, A., Scellato, S., Srinivasan, P., Chen, M., Koverkathu, V., Dalibard, V., Xu, Y., Saeta, B., Anderson, K., Sellam, T., Fernando, N., Huot, F., Jung, J., Varadarajan, M., Quinn, M., Raul, A., Le, M., Habalov, R., Clark, J., Jalan, K., Bullard, K., Singhal, A., Luong, T., Wang, B., Rajayogam, S., Eisen-schlos, J., Jia, J., Finchelstein, D., Yakubovich, A., Balle, D., Fink, M., Agarwal, S., Li, J., Dvijotham, D., Pal, S., Kang, K., Konzelmann, J., Beattie, J., Dousse, O., Wu, D., Crocker, R., Elkind, C., Jonnalagadda, S. R., Lee, J., Holtmann-Rice, D., Kallarackal, K., Liu, R., Vnukov, D., Vats, N., Invernizzi, L., Jafari, M., Zhou, H., Taylor, L., Prendki, J., Wu, M., Eccles, T., Liu, T., Kopparapu, K., Beaufays, F., Angermueller, C., Marzoca, A., Sarcar, S., Dib, H., Stanway, J., Perbet, F., Trdin, N., Sterneck, R., Khorlin, A., Li, D., Wu, X., Goenka, S., Madras, D., Goldshtain, S., Gierke, W., Zhou, T., Liu, Y., Liang, Y., White, A., Li, Y., Singh, S., Bahargam, S., Epstein, M., Basu, S., Lao, L., Ozturel, A., Crous, C., Zhai, A., Lu, H., Tung, Z., Gaur, N., Walton, A., Dixon, L., Zhang, M., Globerson, A., Uy, G., Bolt, A., Wiles, O., Nasr, M., Shumailov, I., Selvi, M., Piccinno, F., Aguilar, R., McCarthy, S., Khalman, M., Shukla, M., Galic, V., Carpenter, J., Villela, K., Zhang, H., Richardson, H., Martens, J., Bosnjak, M., Belle, S. R., Seibert, J., Alnahlawi, M., McWilliams, B., Singh, S., Louis, A., Ding, W., Popovici, D., Simicich, L., Knight, L., Mehta, P., Gupta, N., Shi, C., Fatehi, S., Mitrovic, J., Grills, A., Pagadora, J., Petrova, D., Eisen-bud, D., Zhang, Z., Yates, D., Mittal, B., Tripuraneni, N., Assael, Y., Brovelli, T., Jain, P., Velimirovic, M., Akbulut, C., Mu, J., Macherey, W., Kumar, R., Xu, J., Qureshi, H., Comanici, G., Wiesner, J., Gong, Z., Ruddock, A., Bauer, M., Felt, N., GP, A., Arnab, A., Zelle, D., Rothfuss, J., Rosgen, B., Shenoy, A., Seybold, B., Li, X., Mudigonda, J., Erdogan, G., Xia, J., Sims, J., Michi, A., Yao, Y., Yew, C., Kan, S., Caswell, I., Radebaugh, C., Elisseeff, A., Valenzuela, P., McKinney, K., Paterson, K., Cui, A., Latorre-Chimoto, E., Kim, S., Zeng, W., Durden, K., Ponnapalli, P., Sosea, T., Choquette-Choo, C. A., Manyika, J., Robenek, B., Vashisht, H., Pereira, S., Lam, H., Velic, M., Owusu-Afriyie, D., Lee, K., Bolukbasi, T., Parrish, A., Lu, S., Park, J., Venkatraman, B., Talbert, A., Rosique, L., Cheng, Y., Sozanschi, A., Paszke, A., Kumar, P., Austin, J., Li, L., Salama, K., Kim, W., Dukkipati, N., Baryshnikov, A., Kaplanis, C., Sheng, X., Chervonyi, Y., Unlu, C., de Las Casas, D., Askham, H., Tunyasuvunakool, K., Gimeno, F., Poder, S., Kwak, C., Miecnikowski, M., Mirrokni, V., Dimitriev, A., Parisi, A., Liu, D., Tsai, T., Shevlane, T., Kouridi, C., Garmon, D., Goedeckemeyer, A., Brown, A. R., Vijayakumar, A., Elqursh, A., Jazayeri, S., Huang, J., Carthy, S. M., Hoover, J., Kim, L., Kumar, S., Chen, W., Biles, C., Bingham, G., Rosen, E., Wang, L., Tan, Q., Engel, D., Pongetti, F., de Cesare, D., Hwang, D., Yu, L., Pullman, J., Narayanan, S., Levin, K., Gopal, S., Li, M., Aharoni, A., Trinh, T., Lo, J., Casagrande, N., Vij, R., Matthey, L., Ramadhana, B., Matthews, A., Carey, C., Johnson, M., Goranova, K., Shah, R., Ashraf, S., Dasgupta, K., Larsen, R., Wang, Y., Vuyyuru, M. R., Jiang, C., Ijazi, J., Osawa, K., Smith, C., Boppana, R. S., Bilal, T., Koizumi, Y., Xu, Y., Altun, Y., Shabat, N., Bariach, B., Korchemniy, A., Choo, K., Ronneberger, O., Iwuanyanwu, C., Zhao, S., Soergel, D., Hsieh, C.-J., Cai, I., Iqbal, S., Sundermeyer, M., Chen, Z., Bursztein, E., Malaviya, C., Biadsy, F., Shroff, P., Dhillon, I., Latkar, T., Dyer, C., Forbes, H., Nicosia, M., Nikolaev, V., Greene, S., Georgiev, M., Wang, P., Martin, N., Sedghi, H., Zhang, J., Banzal, P., Fritz, D., Rao, V., Wang, X., Zhang, J., Patraucean, V., Du, D., Mordatch, I., Jurin, I., Liu, L., Dubey, A., Mohan, A., Nowakowski, J., Ion, V.-D., Wei, N., Tojo, R., Raad, M. A., Hudson, D. A., Keshava, V., Agrawal, S., Ramirez, K., Wu, Z., Nguyen, H., Liu, J., Sewak, M., Petrini, B., Choi, D., Philips, I., Wang, Z., Bica, I., Garg, A., Wilkiewicz, J., Agrawal, P., Li, X., Guo, D., Xue, E., Shaik, N., Leach, A., Khan, S. M., Wiesinger, J., Jerome, S., Chakladar, A., Wang, A. W., Ornduff, T., Abu, F., Ghaffarkhah, A., Wainwright, M., Cortes, M., Liu, F., Maynez, J., Terzis, A., Samangouei, P., Mansour, R., Kepa, T., Aubet, F.-X., Algymr, A., Banica, D., Weisz, A., Orban, A., Senges, A., Andrejczuk, E., Geller, M., Santo, N. D., Anklin, V., Merey, M. A., Baeuml, M., Strohman, T., Bai, J., Petrov, S., Wu, Y., Hassabis, D., Kavukcuoglu, K., Dean, J., and Vinyals, O. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL <https://arxiv.org/abs/2403.05530>.
- Gemma Team, Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.-B., Neyshabur, B., Bachem, O., Walton, A., Severyn, A., Parish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Pater-son, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C., Choquette-Choo, C. A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozińska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucińska, H., Batra, H., Dhand, H., Nar-

- dini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J. P., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J., Gordon, J., Lipschultz, J., Newlan, J., yeong Ji, J., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L. L., Usui, L., Sifre, L., Heuermann, L., Lago, L., McNealus, L., Soares, L. B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Görner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R. A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S. M., Perrin, S., Arnold, S. M. R., Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kociský, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R., and Andreev, A. Gemma 2: Improving open language models at a practical size, 2024. URL <https://arxiv.org/abs/2408.00118>.
- Ghazvininejad, M., Gonen, H., and Zettlemoyer, L. Dictionary-based phrase-level prompting of large language models for machine translation, 2023. URL <https://arxiv.org/abs/2302.07856>.
- Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzmán, F., and Fan, A. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl_a_00474. URL <https://aclanthology.org/2022.tacl-1.30>.
- Guerreiro, N. M., Rei, R., Stigt, D. v., Coheur, L., Colombo, P., and Martins, A. F. T. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024. doi: 10.1162/tacl_a_00683. URL <https://aclanthology.org/2024.tacl-1.54>.
- He, Z., Liang, T., Jiao, W., Zhang, Z., Yang, Y., Wang, R., Tu, Z., Shi, S., and Wang, X. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 11:229–246, 2024.
- Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., and Awadalla, H. H. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*, 2023.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models, 2022. URL <https://arxiv.org/abs/2203.15556>.
- ImaniGooghari, A., Lin, P., Kargaran, A. H., Severini, S., Jalili Sabet, M., Kassner, N., Ma, C., Schmid, H., Martins, A., Yvon, F., and Schütze, H. Glot500: Scaling multilingual corpora and language models to 500 languages. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1082–1117, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.61. URL <https://aclanthology.org/2023.acl-long.61>.
- Juraska, J., Finkelstein, M., Deutsch, D., Siddhant, A., Mirzazadeh, M., and Freitag, M. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C. (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 756–767, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.63. URL <https://aclanthology.org/2023.wmt-1.63>.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models, 2020. URL <https://arxiv.org/abs/2001.08361>.
- Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., and Sabharwal, A. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_nGgzQjzaRy.
- Ki, D. and Carpuat, M. Guiding large language models to post-edit machine translation with error annotations. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Findings of the Association for Computational Linguistics*, 2023.

- Linguistics: NAACL 2024*, pp. 4253–4273, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.265. URL <https://aclanthology.org/2024.findings-naacl.265>.
- Koehn, P. Statistical significance tests for machine translation evaluation. In Lin, D. and Wu, D. (eds.), *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-3250>.
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 22199–22213. Curran Associates, Inc., 2022.
- Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- Lai, W., Mesgar, M., and Fraser, A. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. In Ku, L.-W., Martins, A., and Srikanth, V. (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 8186–8213, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.488. URL <https://aclanthology.org/2024.findings-acl.488>.
- Lu, H., Yang, H., Huang, H., Zhang, D., Lam, W., and Wei, F. Chain-of-dictionary prompting elicits translation in large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 958–976, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.emnlp-main.55>.
- Lù, X. H. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring, 2024. URL <https://arxiv.org/abs/2407.03618>.
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., and Clark, P. Self-refine: Iterative refinement with self-feedback. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594. Curran Associates, Inc., 2023.
- Moslem, Y., Haque, R., Kelleher, J. D., and Way, A. Adaptive machine translation with large language models. In Nurminen, M., Brenner, J., Koponen, M., Latomaa, S., Mikhailov, M., Schierl, F., Ranasinghe, T., Vanmassenhove, E., Vidal, S. A., Aranberri, N., Nunziatini, M., Escartín, C. P., Forcada, M., Popovic, M., Scarton, C., and Moniz, H. (eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 227–237, Tampere, Finland, June 2023. European Association for Machine Translation. URL <https://aclanthology.org/2023.eamt-1.22>.
- Mu, Y., Reheman, A., Cao, Z., Fan, Y., Li, B., Li, Y., Xiao, T., Zhang, C., and Zhu, J. Augmenting large language model translators via translation memories. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10287–10299, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.653. URL <https://aclanthology.org/2023.findings-acl.653>.
- Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Le Scao, T., Bari, M. S., Shen, S., Yong, Z. X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raffel, C. Crosslingual generalization through multitask finetuning. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.891. URL <https://aclanthology.org/2023.acl-long.891>.
- Niklaus, C., Freitas, A., and Handschuh, S. MinWik-iSplit: A sentence splitting corpus with minimal propositions. In van Deemter, K., Lin, C., and Takamura, H. (eds.), *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 118–123, Tokyo, Japan, October–November 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-8615. URL <https://aclanthology.org/W19-8615>.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pp. 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://doi.org/10.3115/1073083.1073135>.

- Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., and Tao, D. Towards making the most of ChatGPT for machine translation. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5622–5633, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.373. URL <https://aclanthology.org/2023.findings-emnlp.373>.
- Popović, M. chrF: character n-gram F-score for automatic MT evaluation. In Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Hokamp, C., Huck, M., Logacheva, V., and Pecina, P. (eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL <https://aclanthology.org/W15-3049>.
- Popović, M. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pp. 612–618, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/W17-4770>.
- Post, M. A call for clarity in reporting BLEU scores. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K. (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL <https://aclanthology.org/W18-6319>.
- Puduppully, R., Kunchukuttan, A., Dabre, R., Aw, A. T., and Chen, N. DecoMT: Decomposed prompting for machine translation between related languages using large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4586–4602, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.279. URL <https://aclanthology.org/2023.emnlp-main.279>.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. Unbabel’s participation in the WMT20 metrics shared task. In Barrault, L., Bojar, O., Bougares, F., Chatterjee, R., Costa-jussà, M. R., Federmann, C., Fishel, M., Fraser, A., Graham, Y., Guzman, P., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Morishita, M., Monz, C., Nagata, M., Nakazawa, T., and Negri, M. (eds.), *Proceedings of the Fifth Conference on Machine Translation*, pp. 911–920, Online, November 2020. Association for Computational Linguistics. URL <https://aclanthology.org/2020.wmt-1.101>.
- Rei, R., Guerreiro, N. M., Pombal, J., van Stigt, D., Treviso, M., Coheur, L., de Souza, J. G., and Martins, A. F. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. *arXiv preprint arXiv:2309.11925*, 2023.
- Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M. M., and Gatford, M. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pp. 109–126. Gaithersburg, MD: NIST, January 1995. URL <https://www.microsoft.com/en-us/research/publication/okapi-at-trec-3/>.
- Schwenk, H., Wenzek, G., Edunov, S., Grave, E., Joulin, A., and Fan, A. CCMMatrix: Mining billions of high-quality parallel sentences on the web. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *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)*, pp. 6490–6500, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.507. URL <https://aclanthology.org/2021.acl-long.507>.
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. 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, 2024. Curran Associates Inc.
- Singh, S., Vargas, F., D’souza, D., Karlsson, B., Mahendiran, A., Ko, W.-Y., Shandilya, H., Patel, J., Mataciunas, D., O’Mahony, L., Zhang, M., Hettiarachchi, R., Wilson, J., Machado, M., Moura, L., Krzemiński, D., Fadaei, H., Ergun, I., Okoh, I., Alaagib, A., Mudannayake, O., Alyafeai, Z., Chien, V., Ruder, S., Guthikonda, S., Alghamdi, E., Gehrmann, S., Muennighoff, N., Bartolo, M., Kreutzer, J., Üstün, A., Fadaee, M., and Hooker, S. Aya dataset: An open-access collection for multilingual instruction tuning. In Ku, L.-W., Martins, A., and Srikanth, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11521–11567, Bangkok,

- Thailand, August 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.acl-long.620>.
- Tanzer, G., Suzgun, M., Visser, E., Jurafsky, D., and Melas-Kyriazi, L. A benchmark for learning to translate a new language from one grammar book. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=tbVWug9f2h>.
- Turcato, D. and Popowich, F. What is example-based machine translation? In Carl, M. and Way, A. (eds.), *Workshop on Example-Based machine Translation*, Santiago de Compostela, Spain, September 18-22 2001. URL <https://aclanthology.org/2001.mtsummit-ebmt.7>.
- Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., and Foster, G. F. Prompting palm for translation: Assessing strategies and performance. In Rogers, A., Boyd-Graber, J. L., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pp. 15406–15427. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.859. URL <https://doi.org/10.18653/v1/2023.acl-long.859>.
- Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=1PL1NIMMrw>.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022.
- Xu, H., Kim, Y. J., Sharaf, A., and Awadalla, H. H. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*, 2024a. URL <https://openreview.net/forum?id=farT6XXntP>.
- Xu, H., Murray, K., Koehn, P., Hoang, H., Eriguchi, A., and Khayrallah, H. X-alma: Plug & play modules and adaptive rejection for quality translation at scale, 2024b. URL <https://arxiv.org/abs/2410.03115>.
- Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Durme, B. V., Murray, K., and Kim, Y. J. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Forty-first International Conference on Machine Learning*, 2024c. URL <https://openreview.net/forum?id=5liwkioZpn>.
- Xu, W., Deutsch, D., Finkelstein, M., Juraska, J., Zhang, B., Liu, Z., Wang, W. Y., Li, L., and Freitag, M. LLM-Refine: Pinpointing and refining large language models via fine-grained actionable feedback. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1429–1445, Mexico City, Mexico, June 2024d. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.92. URL <https://aclanthology.org/2024.findings-naacl.92>.
- Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. 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*, pp. 483–498, 2021.
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., and Narasimhan, K. Tree of thoughts: Deliberate problem solving with large language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822. Curran Associates, Inc., 2023.
- Yasunaga, M., Chen, X., Li, Y., Pasupat, P., Leskovec, J., Liang, P., Chi, E. H., and Zhou, D. Large language models as analogical reasoners. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=AgDICX1h50>.
- Zebaze, A., Sagot, B., and Bawden, R. In-context example selection via similarity search improves low-resource machine translation, 2024a. URL <https://arxiv.org/abs/2408.00397>.
- Zebaze, A. R., Sagot, B., and Bawden, R. Tree of problems: Improving structured problem solving with compositionality. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18028–18047, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. URL <https://aclanthology.org/2024.emnlp-main.1001>.
- Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., Tam, W. L., Ma, Z., Xue,

Y., Zhai, J., Chen, W., Liu, Z., Zhang, P., Dong, Y., and Tang, J. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=-Aw0rrrPUF>.

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

Zhang, Z., Zhang, A., Li, M., and Smola, A. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*, 2023b. URL <https://openreview.net/forum?id=5NTt8GFjUHkr>.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q. V., and Chi, E. H. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=WZH7099tgfM>.

Zhu, W., Liu, H., Dong, Q., Xu, J., Huang, S., Kong, L., Chen, J., and Li, L. Multilingual machine translation with large language models: Empirical results and analysis. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2765–2781, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.176. URL <https://aclanthology.org/2024.findings-naacl.176>.

Üstün, A., Aryabumi, V., Yong, Z.-X., Ko, W.-Y., D'souza, D., Onilude, G., Bhandari, N., Singh, S., Ooi, H.-L., Kayid, A., Vargas, F., Blunsom, P., Longpre, S., Muenighoff, N., Fadaee, M., Kreutzer, J., and Hooker, S. Aya model: An instruction finetuned open-access multilingual language model, 2024. URL <https://arxiv.org/abs/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 & 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 10, 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.

Table 10. Full quantitative BLEU and chrF++ results for English→Nko on FLORES 200 Nko’s split derived by Doumbouya et al. (2023).

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

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 11. 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.

Table 11. Comparison between the ensembling strategy and each of its components (MetricX scores).

	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

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 Translate⁸. We report the results in Table 12. 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.

⁸<https://translate.google.com/>

Table 12. Full MetricX results for ten French→X directions from FLORES 200.

	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

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 13, CompTra fails to outperform few-shot MT. We 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.

Table 13. Full MetricX results for five English→X high-resource 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

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 1 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 14 we observe that CompTra is still the best strategy, performing better than the others 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 15, 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 translates 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.

Table 14. 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).

Methods	Amharic		Burmese		Fijian		Khmer		Lao	
	XCOMET	MetricX	COMET	MetricX	COMET	MetricX	COMET	MetricX	XCOMET	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	
	XCOMET	MetricX	COMET	MetricX	COMET	MetricX	COMET	MetricX	XCOMET	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

B. Additional Results

B.1. BLEU and chrF++ results on FLORES 200

As mentioned previously we additionally present results with BLEU and chrF++ scores of CompTra against baselines in Table 16) for transparency reasons. 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 17, 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.

B.2. BLEU and chrF++ results on NTREX 128

We present results with BLEU and chrF++ scores of CompTra against baselines on NTREX 128 in Table 18 and TICO-19 in Table 19. The results are the same as in Table 2 and Table 3 when CompTra outperforms few-shot MT with BM25 and SONAR.

B.3. BLEU and chrF++ results on FLORES 200 with small LMs

We reported that CompTra works very well with smaller LMs in Table 5 by reporting the MetricX scores. In Table 20, 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 21. 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.

Table 15. Full MetricX-QE results for ten English→ X directions. We compare how accurately phrases are translated compared to main sentences.

	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 16. 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++								
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++								
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). 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++								
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++								
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 NTREX 128 (Federmann et al., 2022; Barrault et al., 2019).

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 19. 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++								
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 20. 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.

Methods	Amharic		Burmese		Fijian		Khmer		Lao	
	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++								
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 21. 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 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

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 & Yvon, 2023; Zebaze et al., 2024a). It occurs when translating into LRLs with base models but also with instruction 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. For the statistical significant comparison between each pair strategies, we follow (Koehn, 2004) and use paired bootstrap resampling with 300 samples of 500 sentences and a p -value threshold of 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 22, we list the links to the relevant resources used for experiments.

Table 22. Links to datasets, benchmarks and models.

<i>Datasets</i>	
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	https://huggingface.co/datasets/mteb/NTREX
TICO-19	https://huggingface.co/datasets/gmnlp/tico19
<i>Models evaluated</i>	
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-quantz/Meta-Llama-3.1-70B-Instruct-AWQ-INT4
NLLB-200-distilled-600M	https://huggingface.co/facebook/nllb-200-distilled-600M
<i>Other resources</i>	
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/xhluga/bm25s

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. 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

Table 23. LLaMA 3.1 70B It's generations in Samoan.

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 fa'amatelina o le fa'amaonia, o lea e fa'aigoa ai o ia i totolu o le laupesi o tagata fa'atau tupe. E to'aitiiti tagata e fa'aaluina o latou galuega i le fa'atauina 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 totolu o tagata fa'afogaga pepa. E iai ni tagata e fa'afogaga i le galuega a ia e tasi.
Zero-shot + Refine	O le galuega a ia o le aofaiga ma le fa'amatelina o le fa'amaonia, o lea e fa'aigoa ai o ia i totolu o le laupesi o tagata fa'atau tupe. E to'aitiiti tagata e fa'aaluina o latou galuega i le fa'atauina o ana galuega e tasi.
SBYS	O le galuega a ia o le aofaiga ma le fa'amatelina o le tulaga e fa'amaonia, o le mea lea e fa'aigoa ai o ia i totolu o le au fa'atau tupe, e le'i tele ni tagata e fa'atau tupe e fa'aalu i le galuega a ia.
MAPS	O lana galuega e fa'amaonia lelei ma le fa'amatelina o le aiga o le igoa o le aiga i totolu o tagata fa'atau pese. E to'aitiiti tagata e fa'apipi'i i le fa'atauina o lana galuega e tasi.
TEaR	O lana galuega e fa'amaonia lelei ma le fa'amatelina, o le mea lea e fa'aalu ai o ia i totolu o le vae to'aitiiti o "igoa o le aiga" i totolu o tagata fa'atau tupe. E to'aitiiti e fa'apipi'i i le fa'atauina o lana galuega e aunoa.
5-shot BM25	O lana galuega e fa'amaonia lelei ma le fa'amatelina o le aofa'i, o le mea lea e fa'aigoa ai o ia o se tasi o le vae to'aitiiti o "igoa o fale" i totolu o tagata fa'atau tupe. E to'aitiiti e fa'aoga i le fa'atauina 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 totolu o tagata fa'afogae. E iai nisi e fa'aaluina i le fa'atauina o lana galuega e tasi.
5-shot BM25 + Refine	O lana galuega e fa'amaonia lelei ma le fa'amatelina o le aofa'i, o le mea lea e fa'aigoa ai o ia o se tasi o tagata e sili ona lauiloa i totolu o tagata fa'atau tupe. E to'aitiiti e fa'aoga i le fa'atauina o lana galuega e tasi.
CompTra	O lana galuega e fa'amaonia lelei ma le fa'afouina, o se tasi o ni vaaiaga lauiloa i totolu o tagata fa'aalu i tupe. E i ai nisi e sili ona fa'aaaua i le fa'asoia 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.