# In-Context Example Selection via Similarity Search Improves Low-Resource Machine Translation

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

## Abstract

The ability of generative large language models (LLMs) to perform in-context learning has given rise to a large body of research into how best to prompt models for various natural language processing tasks. In this paper, we focus on machine translation (MT), a task that has been shown to benefit from in-context translation examples. However no systematic studies have been published on how best to select examples, and mixed results have been reported on the usefulness of similarity-based selection over random selection, although these results have mainly been shown for high-resource languages only. We provide a study covering multiple LLMs and in-context example retrieval strategies. Contrarily to previously published results, we find that retrieval based on sentence embedding similarity can improve MT, especially for low-resource language directions, and we also discuss the balance between selection pool diversity and quality.[1]

## 1 Introduction

In-context learning (ICL; Brown et al., 2020) for large language models (LLMs) has proved successful for various tasks, including machine translation (MT) (Bawden and Yvon, 2023; Zhang et al., 2023a; Zhu et al., 2023; Mu et al., 2023; Jiao et al., 2023; Moslem et al., 2023; Xu et al., 2024; Lyu et al., 2024; Lai et al., 2024). Usually, in-context examples for MT are randomly sampled from a parallel corpus. However, existing work in question answering (Liu et al., 2022) and text classification (Zhao et al., 2021) has shown that the choice of in-context examples considerably influences ICL outcomes. This aspect has been explored in MT through example retrieval via similarity search, where in-context examples are chosen based on their similarity to the sentence to be translated. However, consensus on its efficacy has not been reached. Vilar et al. (2023) found that retrieving similar sentences does not yield more benefits than selecting them randomly when the selection pool contains only high-quality samples. Their experiments focused on high-resource directions. Zhu et al. (2023) and Hendy et al. (2023) arrived at the same conclusion when examining other high-resource directions. However, Agrawal et al. (2023) surpassed the random baseline by using examples retrieved with BM25 and further improved performance through a re-ranking procedure. Zhang et al. (2023a) observed a correlation between the use of similar examples and performance but cautioned that the correlation may not be strong enough. Not only do these mixed results show that it is not clear whether example selection can provide gains, but the impact of few-shot example selection for low-resource languages remains underexplored. Existing research also often overlooks the impact of the size and quality of the selection pool, and there is a lack of analysis across LLMs of different scales.

In this work, we aim to address these gaps by systematically analyzing example retrieval via similarity search. We benchmark multiple similarity metrics based on multilingual sentence embeddings across various open-access LLMs. We consider translations from English to French, German, Swahili and Wolof to account for different levels of resourcedness. We compare the use of sentence embeddings and existing approaches, and we assess the robustness of this strategy against different selection pool compositions when translating from English to Swahili. Additionally, we highlight potential problems with the evaluation of LLM-based MT and propose a more appropriate evaluation protocol. Our analysis suggests that example retrieval via similarity search only marginally improves MT over random sampling for high-resource languages. However, for the first time, we observe significant gains across all metrics when translating into low-

[1] Code and outputs are available at https://github.com/ArmelRandy/ICL-MT.



*Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252
April 29 - May 4, 2025 

resource languages. These results are observable across LLMs of multiple scales.

## 2 Background and Related Work

**In-Context Learning (ICL).** After Brown et al. (2020) demonstrated GPT-3's strong zero-shot and few-shot abilities on language understanding benchmarks, a lot of effort was put into empirically analyzing ICL. Zhao et al. (2021) showed that the prompt format, the quality of the examples and their order all have an effect on performance, although it has been shown, for example by Min et al. (2022) for few-shot text classification, that performance can plateau as the number of examples included increases. Another line of work explored the design of prompting strategies with most results obtained on reasoning tasks: chain of thought (Wei et al., 2022; Kojima et al., 2022; Zhang et al., 2023b), self-consistency (Wang et al., 2023; Chen et al., 2023) and hierarchical approaches (Yao et al., 2023; Besta et al., 2024; Zebaze et al., 2024).

**Using LLMs for Machine Translation.** In MT, comparing LLMs and understanding their behaviour in few-shot settings has motivated multiple studies. Lin et al. (2022) showed that XGLM 7.5B outperforms GPT-3 6.7B in 32-shot for multiple translation directions. Vilar et al. (2023) used PALM (Chowdhery et al., 2022) for few-shot MT. They ran experiments on high resource languages and concluded that the quality of the selection pool has a high impact on few-shot MT. Zhang et al. (2023a) and Bawden and Yvon (2023) respectively analyzed GLM-130B (Zeng et al., 2023) and BLOOM (BigScience Workshop et al., 2023) for few-shot MT. They both highlighted the importance of the prompt format inter alia. Hendy et al. (2023) demonstrated the competitiveness of GPT models prompted in few-shot against commercial MT systems. Most of these works focus on high-resource languages, but Hendy et al. (2023) used two low-resource languages (Hausa and Icelandic) to demonstrate that GPT models lag behind the best MT systems and Bawden and Yvon (2023) studied 1-shot MT between low-resource languages pairs. Zhu et al. (2023) conducted a systematic study in which they compared eight LLMs for few-shot MT in 102 languages covering different resource levels, although most of their experiments were done with eight randomly picked few-shot examples.

**Similarity search for example selection.** While a majority of works, including those in MT, use few-shot examples that are randomly selected, others explore how selecting particular examples can impact performance. This is often achieved by mining sentences similar to the one to be processed, generally based on sentence vector representations based on token-level language models (e.g. RoBERTa; Liu et al., 2019) or on sentence embedding models (e.g. LASER2; Heffernan et al., 2022). Liu et al. (2022) showed that $k$-NN retrieval with fine-tuned RoBERTa models improved GPT-3 performance on question answering and table-to-text generation tasks. Vilar et al. (2023) implemented $k$-NN retrieval with RoBERTa and bag-of-word embeddings for few-shot MT between high-resource language pairs. Similarly, Zhu et al. (2023) compared BM25 (Robertson et al., 1995) to example retrieval with a sentence embedding for MT from English to German and Russian. They both conclude that the use of similar examples is comparable to that of random examples for a high quality selection pool. Hendy et al. (2023) used LaBSE (Feng et al., 2022) to build a high-quality selection pool and/or to perform high-quality example selection. Their experiments on German, Russian and Chinese showed the irrelevance of quality selection from a high quality selection pool. Zhang et al. (2023a) studied the correlation between shot selection and MT performance for multiple strategies including example retrieval with LASER2. Their work mostly focused on Chinese and German for which they reported mixed results. Agrawal et al. (2023) and Bouthors et al. (2024) explored retrieval with BM25 and showed that their re-ranking procedure could improve BLEU scores. The variability in the conclusions regarding the efficacy of similarity-based selection methods highlights the need for a more systematic study covering both high-resource and low-resource languages, which are frequently excluded from these experiments. Moslem et al. (2023) is a notable exception for exploring how ICL with fuzzy matches aids MT from English to Kinyarwanda and other languages.

## 3 Example retrieval via similarity search

Example retrieval via similarity search is a selection strategy for ICL. The idea is to use the input in order to retrieve similar (input, output) pairs from a pool of labeled data, which can then be used as few-shot examples (see Figure 1). It revolves around

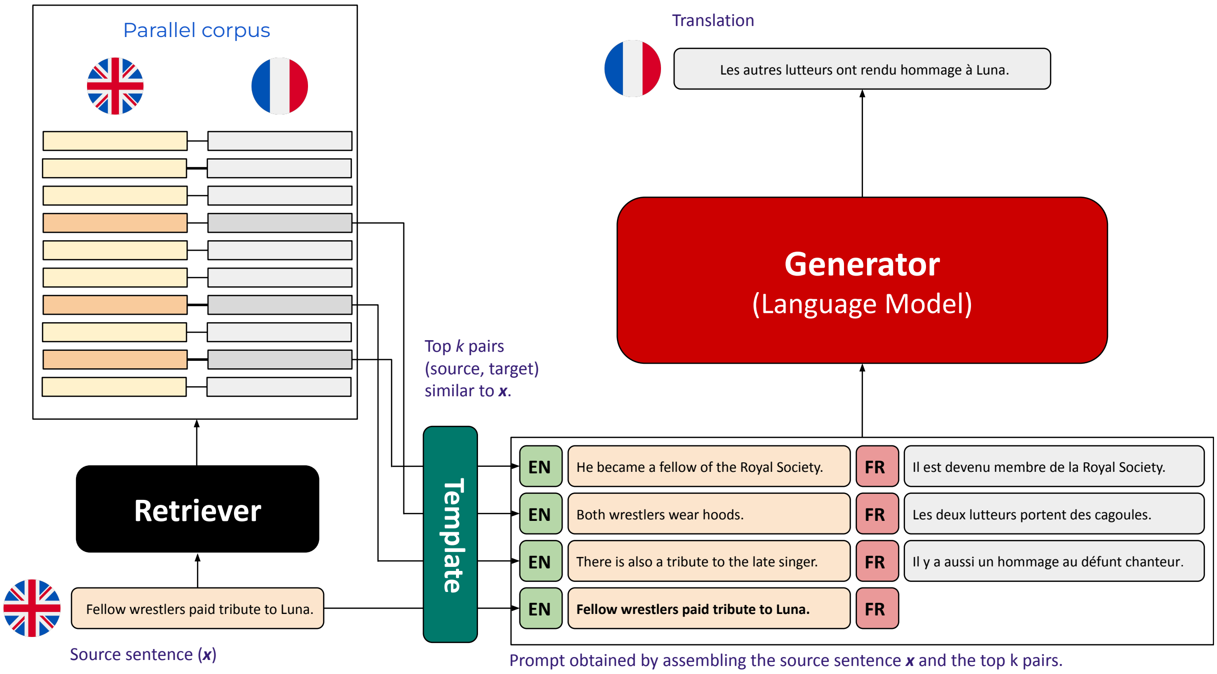


Figure 1: An overview of example retrieval via similarity search for MT. $k$ sentences are first retrieved from the example pool (parallel corpus) based on their similarity to the source sentence. The retrieved sentence pairs are then assembled (as few-shot examples) with the source sentence into a prompt that is fed to a LLM for translation.

the following parameters:

1. **A pool $\mathcal{P}$ from which to retrieve examples for the source sentence $x$.** For MT, the pool corresponds to a set of parallel sentence pairs.
2. **The number $k$ of few-shot examples to retrieve from $\mathcal{P}$.** By definition, $k \leq |\mathcal{P}|$.
3. **A retriever $\mathcal{R}$.** In a similar spirit to RAG (Lewis et al., 2020), its role is to identify similar example pairs to add to the the input prompt. This similarity can be syntactic or semantic. In this work, we model similarity with cosine similarity and we compare this to $n$-gram metrics.
4. **A template to format each example.** This is used to assemble the sentence to translate and the few-shot examples to construct the prompt to be fed to the LLM. By default, the most similar demonstration is the closest to the sentence to translate. We ablate this in Appendix C.2.
5. **An LLM.** The LLM ($p_\theta$) is fed with the prompt in order to obtain the translation. We test a variety of decoder-based LLMs in our study.

In MT, $\mathcal{P}$ consists of the source and target sides of parallel data. Retrieval can be done by analyzing the similarity of the sentence at hand to either the source or target side of each pair in $\mathcal{P}$, which defines two possible approaches, which we refer to as *source-to-source* and *source-to-target*. By default (and unless specified otherwise) we use the *source-to-source* retrieval approach (see Appendix C.8 for the *source-to-target* approach).

## 4 Experimental setup[2]

**Datasets** We work on MT from English (eng) as it is more challenging than translating into English[3] and choose to work with four target languages: two high-resource, French (fra) and German (deu), one mid-resource, Swahili (swa) and one low-resource, Wolof (wol).[4] For evaluation, we mainly use the FLORES-200 (Goyal et al., 2022; Costa-jussà et al., 2022) devtest set containing 1012 examples. We use the FLORES-200 dev set (997 examples) as the selection pool $\mathcal{P}$. We also consider additional examples from the NLLB dataset (Costa-jussà et al., 2022) for experiments involving pool extension. In Appendix C.3, we present additional results on the TICO-19 (Anastasopoulos et al., 2020) benchmark and examine the impact of domain mismatch between the selection pool and the evaluation set on example selection via similarity search.

**Retrievers** We compare five multilingual sentence embeddings: SONAR (Duquenne et al.,

[2]We report implementation details in Appendix A.
[3]See Appendix C.9 for translation into English.
[4]Results on more target languages in Appendix C.7.

2023), Embed v3,[5] E5 (Wang et al., 2022), LaBSE (Feng et al., 2022) and LASER2 (Heffernan et al., 2022). We compare against the following approaches: BM25 (Robertson et al., 1995), R-BM25—consisting in retrieving the top 100 similar candidates with BM25, re-ranking them using the algorithm outlined in (Agrawal et al., 2023) and choosing the $k$ best as demonstrations—, BLEU (Papineni et al., 2002) and RoBERTa-large (Liu et al., 2019) embeddings.[6] We also compare against a random in-pool selection baseline, reporting the average score over three different seeds.

**Models** We test multiple LLMs in our experiments. For reproducibility, we consider state-of-the-art open-access LLMs: BLOOM 7B (Big-Science Workshop et al., 2023), OLMo 7B (Groeneveld et al., 2024), Gemma (2B, 7B) (Gemma Team et al., 2024) LLaMA 2 (7B, 13B and 70B) (Touvron et al., 2023), Mistral 7B v0.1 (Jiang et al., 2023) and Mixtral 8x7B v0.1 (Jiang et al., 2024).

**Evaluation metrics** Historically, BLEU (Papineni et al., 2002) has been the standard MT evaluation metric. The recent advances in deep learning fueled the emergence of neural metrics, one of the most successful being COMET (Rei et al., 2020), which is better correlated with human judgements than BLEU (Rei et al., 2022). Despite this superiority, COMET has some limitations for evaluating MT by LLMs. First, it is inherently limited by the language coverage of its encoder, impairing its reliability for unseen languages (e.g. Wolof). Moreover, it is not robust to the issues of translation in the wrong language and empty translations. Pre-LLM approaches to MT evaluation assumed that MT systems were rarely prone to such issues. However, they have become relevant with the use of LLMs for MT, since these models are not trained for MT specifically, and therefore the premise of a translation being in the correct language does not always hold. The two problems are more likely to appear in zero-shot settings and when few in-context examples are used, especially when prompting a model to generate a low-resource language. We propose to alleviate these issues, later independently identified and investigated by Zouhar et al. (2024) with a simple correction protocol consisting in setting the score of a translation to 0 if it is either empty or written in the wrong target language. We name this variant Language-Aware COMET (laCOMET) which preserves the benefits of COMET while making it robust to the previously mentioned issues[7]. It is worth noting that laCOMET is strictly equivalent to COMET for sentences that do not exhibit the issues that motivated its creation (i.e. non-empty translations in the correct language).

We use laCOMET, based on COMET 22 (Rei et al., 2022) as our main metric. We use fasttext (Bojanowski et al., 2017; Costa-jussà et al., 2022) for language identification, which supports all the languages we work with. For transparency, we also include BLEU calculated using SacreBLEU (Post, 2018)[8] and COMET in the appendix.

**Template** We use the same template across all experiments (see Appendix B.1 for an ablation study).

## 5 Experiments

In Section 5.1 we do a systematic study of example retrieval with several multilingual sentence embeddings for different numbers of in-context examples and families of LLMs, and in Section 5.2 we compare example retrieval with the best performing sentence embedding and the previously mentioned alternative approaches. In Section 5.3 we study the robustness of example retrieval to the size and the diversity of the pool of examples. Finally, in Section 5.4, we focus on English to Swahili and analyze example retrieval for various LLMs at different scales.

### 5.1 Benchmarking of example retrieval with multilingual sentence embeddings

We conduct a benchmarking analysis of example retrieval using multilingual sentence embeddings to evaluate their performance and compare them to random sampling.[9] As demonstrated in Table 1, example retrieval with sentence embeddings consistently outperforms random sampling in few-shot scenarios (up to 10-shot). The performance gain is modest when translating into French and German, typically ranging between 0.1 and 0.5 laCOMET for most LLMs we evaluated, and it tends to narrow as the number of in-context examples increases. However, we note a substantial improvement of around 2.5 in German with BLOOM 7B1. We at-

[5] https://txt.cohere.com/introducing-embed-v3/

[6] More precisely, we use the last hidden state of the first token and send it to the pooling layer.

[7] See Appendix B.2 for more details.

[8] nrefs:1|case:mixed|eff:no|tok:flores200|smooth:exp|version:2.3.2

[9] We provide an analysis of the overlap between their sampling choices in Appendix C.4.

| Method | eng→fra | | | eng→deu | | | eng→swh | | | eng→wol | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| BLOOM 7B1 | | | | | | | | | | | | |
| Embed v3 | 79.6 | 86.7 | **86.7** | 55.2 | **60.1** | **61.0** | 58.6 | **68.4** | 69.4 | 50.4 | 50.2 | 50.7 |
| E5 | **80.4** | 86.6 | **86.7** | 54.5 | **60.1** | 60.6 | **59.8** | 68.2 | 69.3 | **50.9** | **51.4** | 50.7 |
| LaBSE | 79.4 | 86.7 | **86.7** | 55.1 | 59.9 | 60.5 | 58.3 | 67.8 | 69.2 | 49.9 | 51.2 | **52.3** |
| LASER2 | 79.2 | 86.6 | **86.7** | 55.1 | 59.9 | 59.6 | 58.0 | 67.7 | 67.8 | 48.5 | 50.1 | 50.9 |
| SONAR | 79.8 | **86.8** | 86.6 | **55.3** | **60.1** | 60.8 | 57.4 | 68.3 | **69.6** | 50.2 | 50.4 | 51.6 |
| Random | 77.3 | 86.5 | 86.6 | 52.8 | 57.7 | 57.7 | 56.9 | 65.1 | 66.0 | 46.5 | 45.1 | 46.4 |
| Mistral 7B v0.1 | | | | | | | | | | | | |
| Embed v3 | **86.2** | **87.0** | **87.0** | 83.5 | 85.7 | **85.9** | **37.5** | **41.4** | 43.3 | 36.5 | 44.1 | 44.7 |
| E5 | 85.7 | **87.0** | 86.9 | 83.4 | 85.2 | 85.5 | 37.3 | 41.3 | 43.2 | 36.6 | 44.3 | 44.4 |
| LaBSE | 86.2 | 86.7 | **87.0** | 83.3 | 85.3 | 85.6 | 37.0 | 40.1 | 42.3 | **36.7** | 42.6 | 44.6 |
| LASER2 | 86.1 | 86.9 | **87.0** | 83.5 | 85.6 | 85.5 | 35.3 | 38.0 | 40.3 | 32.0 | 42.1 | 43.3 |
| SONAR | 86.1 | 86.9 | **87.0** | **83.6** | **85.8** | **85.9** | 37.2 | 40.6 | **43.5** | 36.4 | **45.0** | **46.1** |
| Random | 85.8 | 86.5 | 86.6 | 83.0 | 85.4 | 85.5 | 32.7 | 33.5 | 33.8 | 26.7 | 33.2 | 36.0 |
| LLaMA 2 7B | | | | | | | | | | | | |
| Embed v3 | 85.8 | 86.1 | 86.3 | 84.0 | 84.9 | 85.0 | **45.7** | **43.7** | **45.6** | 41.8 | 46.2 | **47.1** |
| E5 | 85.8 | **86.2** | **86.4** | **84.1** | 85.2 | 85.2 | 45.1 | 43.3 | 45.3 | **42.3** | **46.5** | 46.9 |
| LaBSE | 85.6 | 86.0 | 86.2 | **84.1** | 85.1 | 85.1 | 44.2 | 42.5 | 44.7 | 40.0 | 43.7 | 45.6 |
| LASER2 | 85.8 | **86.2** | 86.2 | 83.6 | 85.0 | 85.2 | 41.2 | 40.1 | 42.1 | 38.7 | 42.5 | 43.3 |
| SONAR | **85.9** | 86.1 | 86.3 | 83.8 | **85.3** | **85.4** | 45.2 | 43.2 | 45.5 | 39.7 | 45.9 | 46.7 |
| Random | 85.6 | 85.9 | 86.0 | 83.6 | 84.8 | 85.0 | 35.4 | 34.7 | 35.8 | 34.4 | 34.7 | 36.5 |

Table 1: laCOMET results of example retrieval with different sentence embedding methods for $k$-shot settings ($k \in \{1, 5, 10\}$). Underline entries: denote instances where the difference with Random sampling is not significant ($p \geq 0.05$) for at least one of the three seeds.

tribute this greater improvement to the relatively poor performance of BLOOM 7B1 in German as German was not officially included in its training data. For translation into Swahili, the use of sentence embeddings yields gains ranging between 1.7 and 3.4 laCOMET for BLOOM 7B1, 0.6 and 1.6 for Gemma 7B. These gains explode and reach 10 laCOMET when translating into Swahili or Wolof with Mistral 7B v0.1 and LLaMA 2 7B. Furthermore, all sentence embeddings outperform random sampling in a majority of cases. Although there is not a highly significant variation in performance among them, SONAR, Embed v3 and E5 perform slightly better than LaBSE and LASER2 for example retrieval. SONAR yields the best performance with a little advance on Embed v3 and E5. In summary, the use of similar in-context examples yields modest gains for high-resource languages, consistent with previous findings (Zhang et al., 2023a), but we see significant benefits for low-resource languages. We document the same findings in terms of BLEU and COMET in Appendix C.5 and with more LLMs in Appendix C.6.

### 5.2 Comparing to other approaches

We compare the best performing multilingual sentence embeddings model, SONAR against other approaches from the literature in few-shot scenarios. laCOMET scores are given in Table 2.[10] SONAR yields larger performance gains across all directions and LLMs. Following SONAR, BM25 emerges as the second-best approach. Its reliance on $n$-gram-(word-)matching inherently positions it as a strong contender for example selection. However, applying the re-ranking proposed by Agrawal et al. (2023) fails to further improve BM25 in our experimental setup. We attribute this to a lack of diversity in the pool, which hinders its ability to cover each word of the sentences to translate. While RoBERTa achieves performance levels comparable to those of SONAR in French and German, it consistently lags behind in Swahili and Wolof. This discrepancy may be due to RoBERTa not being explicitly trained to produce similar vector representations for similar sentences, resulting in worse choices than SONAR. Nevertheless, it still outper-

[10]We report results with additional LLMs in Appendix C.6.

forms random sampling in our evaluations.

### 5.3 Robustness to the quality and the diversity of the selection pool

The performance of ICL is heavily dependent on the diversity and quality of the selection pool. The initial selection pool is a small set of high quality professional translations. Similar to previous works, we extensively studied example retrieval with a high quality pool. In this set of experiments, we compare the behavior of example retrieval with SONAR and BM25 when translating into Swahili across eight[11] different pool compositions $\mathcal{P}_1, \ldots, \mathcal{P}_8$. Each composition includes samples from FLORES-200 dev set and/or samples from the NLLB dataset (see Section 4). We assess the quality and diversity of each of the eight pool compositions in Table 3 with two key metrics: the Vendi Score (Dan Friedman and Dieng, 2023) and the average perplexity. The Vendi Score, computed with SONAR embeddings, measures diversity, with higher values indicating greater diversity within the composition. The average perplexity, computed using Gemma 2B, measures sample quality, with lower values indicating higher quality samples. In Figure 4, we observe a gradual performance improvement with SONAR and BM25 as the selection pool contains more and more high-quality samples (from $\mathcal{P}_1$ to $\mathcal{P}_4$) in the 5 and 10-shot settings. Although the difference with random sampling is initially modest for both strategies (at $\mathcal{P}_1$), it steadily widens until $\mathcal{P}_4$. The introduction of NLLB samples in the selection pool, which are inherently of lower quality compared to FLORES-200's, induces a decay in the overall quality of outputs for all strategies with random sampling being particularly affected. SONAR emerges as the most robust strategy because it exhibits a lesser performance drop. This motivates the use of example selection via similarity search in scenarios where the quality of the pool is heterogeneous or partially known.

In order to gain more insights into which examples are being selected, we analyze, on average, what is the proportion of in-context examples belonging to the FLORES-200 dev set (i.e. the highest quality examples) among the selected ones. We conduct the analysis in the 10-shot setting with BLOOM 7B1 and report the results in Figure 3. We observe that despite having access to more samples, SONAR is more prone to selecting FLORES's samples than BM25. This suggests that SONAR is better at retrieving more high-quality samples even at the cost of sacrificing the $n$-gram-level similarity to the sentence of interest. This ability to query "good sentences" results in a greater resilience to noisy selection pools. Interestingly, as illustrated in Table 3, the average similarity scores between the retrieved examples in 10-shot increase with the size of the selection pool. This indicates that a larger pool improves the likelihood of retrieving relevant in-context demonstrations, although the quality of the retrieved examples is more important to generate good outputs.

[11]See Appendix C.1 for results with more samples.

### 5.4 Scalability of example retrieval via similarity search

We demonstrate that the advantages of example retrieval are observable across various scales by evaluating it on a range of LLMs with parameter counts ranging from 2B to 70B. Figure 4 highlights the efficacy of example retrieval when translating from English to Swahili. Most LLMs show a performance improvement of at least 4 laCOMET points between the use of SONAR and random sampling for example selection. Interestingly, we observe that even with 20 in-context examples, the gap with random sampling does not plummet; it continues to increase with the number of in-context examples.[12] BM25 consistently outperforms random sampling but does not reach SONAR's laCOMET scores.

## 6 Discussion

**Example selection via similarity search improves MT.** Our results for translation into French and German partially resonate with previous work by Vilar et al. (2023) and Zhu et al. (2023), as we reported a small range of improvement for these languages over random sampling for a high quality pool (between 0.1 and 0.5 laCOMET for most LLMs). However, our experiments on Swahili and Wolof show that example selection can yield significant gains for lower-resource languages. For these languages, and when the LLM's context length allowed it, we did not observe a plateau even at 20-shot as opposed to (Zhu et al., 2023).[13] In addition to a strong performance, example retrieval using SONAR is resilient with lower

[12]OLMo 7B's performance drop in the 20-shot setting is caused by its short context length (2048), which makes most generations empty.

[13]We stopped at 20 because of the limited context length of some of our LLMs (e.g. BLOOM 7B1, OLMo 7B), which would have resulted in truncated contexts and therefore have

| Metric | eng→fra | | | eng→deu | | | eng→swh | | | eng→wol | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| BLOOM 7B1 | | | | | | | | | | | | |
| SONAR | 79.8 | **86.8** | 86.6 | **55.3** | **60.1** | **60.8** | 57.4 | **68.3** | **69.6** | **50.2** | **50.4** | **51.6** |
| BM25 | 78.8 | 86.6 | 86.7 | 54.2 | 59.7 | 59.7 | 57.0 | 66.8 | 68.5 | 49.4 | 49.1 | 50.4 |
| R-BM25 | **82.0** | 86.4 | 86.5 | 52.9 | 57.7 | 58.6 | 54.8 | 64.3 | 65.3 | 42.4 | 43.8 | 45.8 |
| BLEU | 78.2 | 86.7 | 86.6 | 53.6 | 59.2 | 59.9 | 57.0 | 66.2 | 67.4 | 49.5 | 49.5 | 50.9 |
| RoBERTa | 78.5 | 86.7 | **86.8** | 54.1 | 59.3 | 58.4 | **57.9** | 66.0 | 67.1 | 50.0 | 49.4 | 49.9 |
| Random | 77.3 | 86.5 | 86.6 | 52.8 | 57.7 | 57.7 | 56.9 | 65.1 | 66.0 | 46.5 | 45.1 | 46.4 |
| Mistral 7B v0.1 | | | | | | | | | | | | |
| SONAR | 86.1 | **86.9** | **87.0** | **83.6** | **85.8** | **85.9** | **37.2** | **40.6** | **43.5** | **36.4** | **45.0** | **46.1** |
| BM25 | **86.2** | 86.8 | 86.9 | **83.6** | 85.4 | 85.7 | 34.9 | 38.8 | 41.4 | 33.0 | 40.7 | 43.3 |
| R-BM25 | **86.2** | 86.5 | 86.6 | 83.5 | 85.5 | 85.4 | 31.9 | 33.8 | 34.5 | 24.1 | 28.5 | 32.3 |
| BLEU | **86.2** | **86.9** | 86.9 | 83.3 | 85.4 | 85.8 | 35.4 | 37.2 | 39.1 | 32.7 | 40.0 | 42.6 |
| RoBERTa | 85.9 | **86.9** | 86.8 | **83.6** | 85.4 | **85.9** | 33.7 | 35.6 | 37.3 | 32.0 | 39.4 | 42.0 |
| Random | 85.8 | 86.5 | 86.6 | 83.0 | 85.4 | 85.5 | 32.7 | 33.5 | 33.8 | 26.7 | 33.2 | 36.0 |
| LLaMA 2 7B | | | | | | | | | | | | |
| SONAR | **85.9** | 86.1 | **86.3** | **83.8** | **85.3** | **85.4** | **45.2** | **43.2** | **45.5** | **39.7** | **45.9** | **46.7** |
| BM25 | 85.6 | 86.1 | 86.2 | 83.3 | 84.9 | 85.1 | 40.7 | 40.1 | 42.6 | 38.1 | 43.0 | 45.1 |
| R-BM25 | 85.5 | 86.0 | 85.8 | 83.1 | 85.0 | 85.0 | 33.5 | 34.2 | 34.8 | 25.4 | 27.7 | 33.1 |
| BLEU | 85.6 | 86.0 | 86.1 | **83.8** | 85.0 | 85.0 | 38.8 | 39.0 | 40.1 | 36.6 | 41.6 | 43.6 |
| RoBERTa | 85.6 | **86.2** | 86.0 | **83.8** | 85.0 | 85.3 | 39.9 | 38.1 | 39.7 | 38.7 | 42.1 | 43.8 |
| Random | 85.6 | 85.9 | 86.0 | 83.6 | 84.8 | 85.0 | 35.4 | 34.7 | 35.8 | 34.4 | 34.7 | 36.5 |

Table 2: Comparison of example retrieval with SONAR to baseline methods for $k$-shot settings ($k \in \{1, 5, 10\}$). The best performance (laCOMET) for each direction is shown in bold.

| | $\mathcal{P}_1$ | $\mathcal{P}_2$ | $\mathcal{P}_3$ | $\mathcal{P}_4$ | $\mathcal{P}_5$ | $\mathcal{P}_6$ | $\mathcal{P}_7$ | $\mathcal{P}_8$ |
|---|---|---|---|---|---|---|---|---|
| #FLORES samples ($N_1$) | 10 | 100 | 500 | 997 | 997 | 997 | 997 | 997 |
| #NLLB samples ($N_2$) | 0 | 0 | 0 | 0 | 1000 | 5000 | 10000 | 20000 |
| Vendi Score | 9.4 | 81.2 | 274.8 | 388.2 | 384.4 | 349.9 | 347.5 | 349.5 |
| Perplexity | 131.0 | 90.9 | 79.9 | 77.4 | 222.7 | 301.8 | 306.3 | 356.5 |
| BM25 scores | 1.51 | 6.43 | 10.3 | 11.91 | 12.85 | 12.31 | 13.30 | 14.43 |
| SONAR scores | 0.04 | 0.12 | 0.18 | 0.20 | 0.21 | 0.22 | 0.23 | 0.24 |

Table 3: Average average similarity scores between each sentence to be translated and its 10 examples retrieved with SONAR and BM25 for each pool composition.

quality pools, outperforming the random baseline as well as the strong BM25 approach. This robustness is observed for both high- and low-resource directions in terms of BLEU and laCOMET.

**What issues arise when prompting an LLM to translate into a low-resource language?** The zero-shot abilities of LLMs are sensitive to the template, as shown in Appendix B.1. This is caused by two problems. First, there are instances where the model fails to understand the task and generates unrelated outputs (e.g. multiple line breaks, a repetition of the end of the prompt in multiple languages or a continuation of the input sentence). Secondly, there is the inability to accurately perform the task, leading for example to the repetition of the input sentence (potentially with a few modifications), partial translation (e.g. with repeating $n$-grams at the end) and translation in an incorrect language. The first problem is generally minor when we have a good template, a high-resource language and a capable LLM .Moreover, it is mostly solved by using a 1-shot example. This is why there is a huge gap between 0-shot and 1-shot performance as pointed out by Hendy et al. (2023). Low-resource directions would require more shots, typically between 2 and 5. The second problem is more tenacious, particularly for low-resource direc-

a negative impact on scores.

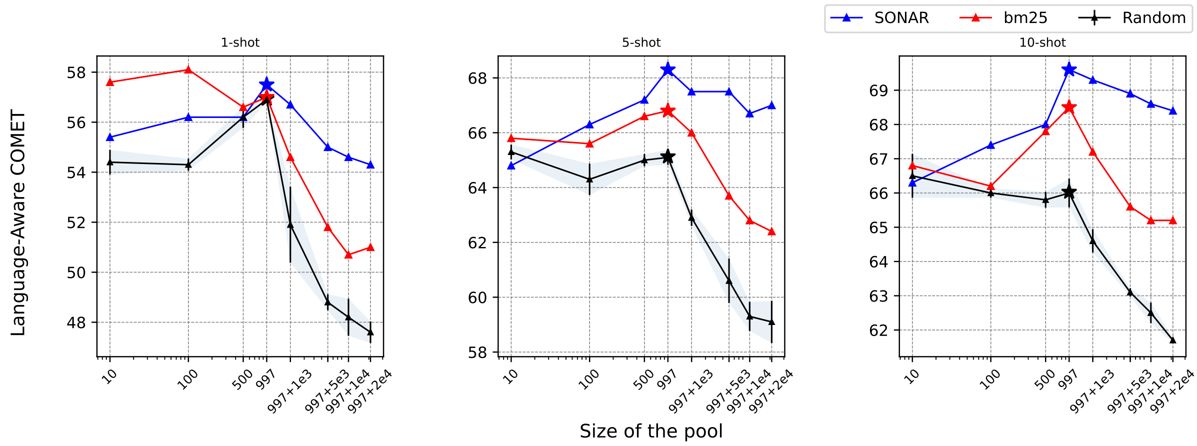


Figure 2: laCOMET scores for example retrieval with SONAR, BM25 and random sampling for various selection pool compositions for eng→swh and BLOOM 7B1. The star indicates the initial pool, i.e. the entire FLORES-200 dev set.

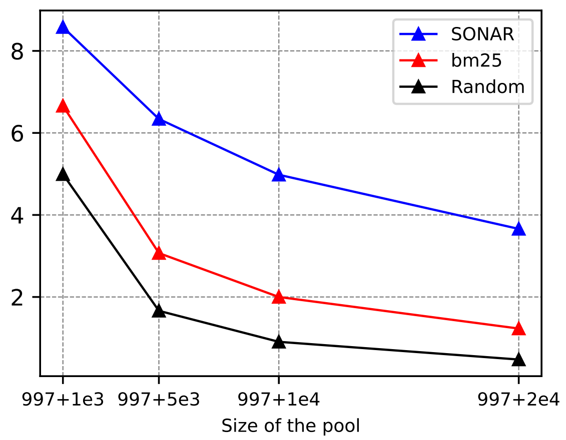


Figure 3: For each pool composition involving FLORES and NLLB samples, the average number of the 10 in-context examples belong to the FLORES-200 dev set when using SONAR, BM25, and random sampling.

tions. As the number of shot increases, the number of translations in the correct language increases and the number of empty translations decreases. However, the scores remain low.

**Why does example selection via similarity search work?** The success of ICL depends on the ability of the LLM to understand the task and its ability to generate a qualitative output given an input. As explained earlier, the task understanding is mostly solved by using few-shot examples. Example selection via similarity search leads to gains in output quality by using qualitative demonstrations aimed at encouraging the LLM to generate higher quality outputs. The impact of example retrieval on the translation from English to French is noticeable at the phrasing level. It makes the LLMs employ different words compared to those used with random sampling to convey the same message. Additionally, it influences the translation of entities (e.g. names of organizations, universities etc.), although we did not observe a consistent pattern in this regard. For translation into Wolof, we observed that example retrieval considerably impacts the rate at which the number of translations in the correct language increases (see Appendix C.10), partially explaining its superior performance. For translation into Swahili, example retrieval helps mitigate the uncontrollable generation of $n$-grams, and its impact on the phrasing is more pronounced than observed for French. The LLMs tend to generate more words in Swahili that are relevant to the context of the sentence to translate.

## 7 Conclusions

We have provided a systematic study of example selection via similarity search as a simple way to improve the MT capabilities of LLMs, comparing the translation quality of multiple open-source LLMs when using a range of different sentence embedding methods to select few-shot examples. We cover four translation directions covering high- and low-resource languages. Our results confirm previous results for high-resource languages that similarity search does not provide significant gains over random sampling. However, we show that the strategy allows LLMs to demonstrate superior translation performance for mid- and low-resource languages. We validated these results across multiple scales of LLMs and example pool sizes. We

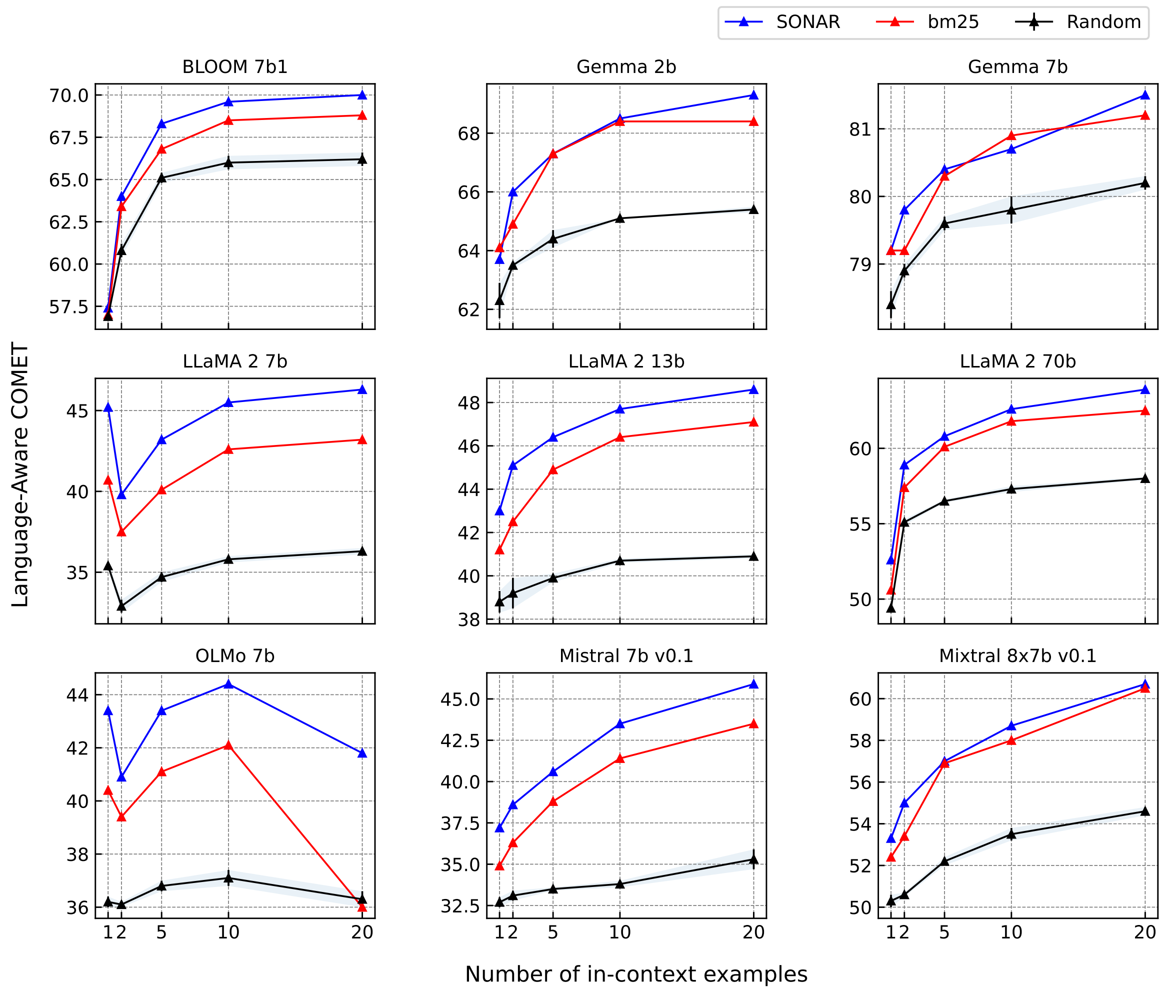


Figure 4: laCOMET scores of example retrieval with SONAR and BM25 compared to random sampling for the $k$-shot setting ($k \in \{1, 2, 5, 10, 20\}$) for eng→swh and nine LLMs. Note that for readability reasons, the Y-axis scales of the figures are not aligned.

also demonstrated that greater diversity in high-quality pools yields better results. Example retrieval is significantly more robust to quality heterogeneity, with sentence embeddings providing the highest resilience.

## Limitations

One inherent limitation of our work is the definition of the concept of similarity; it is a broad and polymorphous concept, and we choose to focus on semantics through the use of sentence embeddings (although it is likely that other aspects are also represented via sentence embeddings). Although other approaches (e.g. more syntax-based) are also possible and would be interesting to explore in future work.

Moreover, despite the gain observed when translating from English to Wolof, it is obvious that most LLMs struggle considerably with this language and other low-resource ones, and this should be a research direction to explore.

## 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. We thank Amadou Dieye for answering our questions about the Wolof language.

## References


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

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

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

Maxime Bouthors, Josep Crego, and François Yvon. 2024. Retrieving examples from memory for retrieval augmented neural machine translation: A systematic comparison. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3022–3039, Mexico City, Mexico. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023. Universal self-consistency for large language model generation.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Dan Dan Friedman and Adji Bousso Dieng. 2023. The vendi score: A diversity evaluation metric for machine learning. *Transactions on machine learning research*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe

Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratan-

chandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 56–60, Vancouver. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

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

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. A paradigm shift: The future of machine translation lies with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Yongyu Mu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. Augmenting large language model translators via translation memories. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10287–10299, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

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

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, and Quoc V Le. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2024. Tree of problems: Improving structured problem solving with compositionality. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18028–18047, Miami, Florida, USA. Association for Computational Linguistics.

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

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

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

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

# A Implementation details

## A.1 Framework and hyperparameters

All our experiments are done with beam search (Freitag and Al-Onaizan, 2017) and a beam size of 2. We use `vLLM` (Kwon et al., 2023) for inference and generate with a maximum sentence length of 100 tokens. In zero-shot settings, we truncate the prediction at the first new line break and ignore any tokens generated afterwards. For the statistical significant comparison with random sampling, we follow (Koehn, 2004) and use paired bootstrap resampling with 300 samples of 500 sentences and a p-value threshold of 0.05.

## A.2 Models

In Table 4, we list the links to the relevant resources used for experiments.

# B Preliminary experiments

## B.1 Template selection

We carried out a preliminary investigation to choose a strong template for our subsequent MT experiments. We compared six potential MT templates (listed in Table 5) in 0-shot and 5-shot settings for three models and the four directions. The BLEU scores are shown in Table 6. The best template for a model does not necessarily work well with another model in the zero-shot setting (e.g. T3 $\geq$ T5 for LLaMA 2 7B but not for Mistral 7B v0.1). We notice that having the end of the prompt written in the target language can dramatically improve zero-shot MT; using template T2 instead of template T1 gives an absolute gain of 11.5 BLEU for BLOOM 7B1, 5.5 for Mistral 7B v0.1 and 0.8 for LLaMA 2 7B for eng$\rightarrow$fra. For eng$\rightarrow$deu, T2 surpasses T1 by 0.2 BLEU for BLOOM 7B1, 4.4 for Mistral 7B v0.1 and 2.7 for LLaMA 2 7B. Similarly, significant gains are observed when using T4 instead of T3. We hypothesize that these improvements can be attributed to the fact that the prompt ending in the target language encourages the model to continue generation in that language, reducing the occurrence of unrelated outputs. The presence of a colon (**:**) at the end of the prompt can have a negative effect on some LLMs such as Mistral 7B v0.1 and LLaMA 2 7B, making them generate dates (with the format YYYY-MM-DD). The performance disparities among templates T1, T2, T5 and T6 disappear in the 5-shot setting but the negative impact of the colon keeps templates T3 and T4 behind. Translating into low-resource languages gives poor zero-shot scores, which prevents a reliable comparison of the templates. The laCOMET scores reported in Table 7 tell the same story and we observe a greater variance between scores in zero-shot compared to BLEU scores. T1, T2, T5, and T6 are the optimal templates for eng$\rightarrow$swh and eng$\rightarrow$wol in few-shot scenarios for all three LLMs. The summary of this analysis is that zero-shot performance varies greatly across templates as observed by Zhang et al. (2023a). This discrepancy tends to disappear in few-shot settings except for adversarial templates. Any of the templates T1, T2, T5 and T6 would allow a fair comparison between models in few-shot scenarios. In this work, we use template T5 because of its simplicity and good few-shot performance.

## B.2 Why laCOMET?

laCOMET was introduced to make COMET robust to instances where LLMs output empty translations or translations in an incorrect target language. Table 8 contains some examples of these issues produced by Mixtral 8x7B v0.1. They often occur when the model has to translate into a language it rarely encountered during its training. In Table 9, these issues give the illusion that Mistral 7B v0.1 outputs better 0-shot Swahili translations than 5-shot ones. Overall, COMET overestimates 0-shot translations and the use of laCOMET reveals this. The laCOMET 0-shot scores are significantly lower than the corresponding COMET scores. As expected, both COMET and laCOMET are close in few-shot settings because the use of in-context demonstrations considerably alleviates these is-

| | |
|---|---|
| *Datasets* | |
| FLORES-200 | `https://huggingface.co/datasets/facebook/flores` |
| NLLB Full dataset | `https://huggingface.co/datasets/allenai/nllb` |
| *Models evaluated* | |
| BLOOM 7B1 | `https://huggingface.co/bigscience/bloom-7b1` |
| OLMo 7B | `https://huggingface.co/allenai/OLMo-7B` |
| Gemma 2B | `https://huggingface.co/google/gemma-2b` |
| Gemma 7B | `https://huggingface.co/google/gemma-7b` |
| LLaMA 2 7B | `https://huggingface.co/meta-llama/Llama-2-7b-hf` |
| LLaMA 2 13B | `https://huggingface.co/meta-llama/Llama-2-13b-hf` |
| LLaMA 2 70B | `https://huggingface.co/TheBloke/Llama-2-70B-AWQ` |
| LLaMA 3 8B | `https://huggingface.co/meta-llama/Meta-Llama-3-8B` |
| Mistral 7B v0.1 | `https://huggingface.co/mistralai/Mistral-7B-v0.1` |
| Mixtral 8x7B v0.1 | `https://huggingface.co/TheBloke/mixtral-8x7B-v0.1-AWQ` |
| RoBERTa | `https://huggingface.co/FacebookAI/roberta-large` |
| *Sentence embeddings* | |
| Cohere | `embed-multilingual-v3.0` |
| E5 | `https://huggingface.co/intfloat/multilingual-e5-large` |
| LaBSE | `https://huggingface.co/sentence-transformers/LaBSE` |
| Laser 2 | `https://github.com/facebookresearch/LASER` |
| SONAR | `https://github.com/facebookresearch/SONAR` |

Table 4: Links to datasets, benchmarks and models.

| ID | Template | Example (eng→fra) |
|---|---|---|
| T1 | [src] ◇ [source] ◇ translates into ◇ [tgt] ◇ | English ◇ I live in Paris. ◇ translates into ◇ French ◇ |
| T2 | $[\text{src}]_{\textbf{src}}$ ◇ [source] ◇ translates into ◇ $[\text{tgt}]_{\textbf{tgt}}$ ◇ | English ◇ I live in Paris. ◇ translates into ◇ Français ◇ |
| T3 | [src]: [source] ◇ [tgt]: | English: I live in Paris. ◇ French: |
| T4 | $[\text{src}]_{\textbf{src}}$: [source] ◇ $[\text{tgt}]_{\textbf{tgt}}$: | English: I live in Paris. ◇ Français: |
| T5 | [src sentence] ◇ [source] ◇ [tgt translation] ◇ | English sentence ◇ I live in Paris. ◇ French translation ◇ |
| T6 | $[\text{src sentence}]_{\textbf{src}}$ ◇ [source] ◇ $[\text{tgt translation}]_{\textbf{tgt}}$ ◇ | English sentence ◇ I live in Paris. ◇ Traduction en français ◇ |

Table 5: Templates considered for template selection. *src* represents the source language (e.g. English), *tgt* the target language (e.g. French) and *source* the sentence to translate. The presence of the subscripts **src** and **tgt** indicates that the words are written in the source language and the target language, respectively. ◇ represents a line break.

sues. Moreover, we evaluated BLOOM 7B1's zero- and 5-shot translations across the 6 templates with Metric-X23 Large (Juraska et al., 2023) and reported the results in Table 10. Despite being larger than COMET (in terms of number of parameters), it struggles to differentiate between zero-shot and few-shot Wolof translations. It also considers BLOOM 7B's zero-shot Swahili translations with Template T5, which are mostly in English, to be close in quality with the 5-shot translations. laCOMET is more robust to the pitfalls of in-context MT than COMET and Metric-X23. A language-aware version of Metric-X23 would have the same advantages as laCOMET but would be more expensive to use.

# C Additional results

## C.1 Impact of scaling the selection pool

As discussed in Section 5.3, we showed that example selection via similarity search is crucial not only when the selection pool contains high-quality samples but also when it is heterogeneous with many low-quality samples. The performance drop observed after introducing NLLB samples does not undermine the importance of similarity search; rather, it highlights that poor in-context examples can negatively impact translation performance. However, similarity search remains highly robust in such scenarios. In Figure 2, we reported results with up to 20K NLLB samples, while Table 11 shows that even with 1M NLLB samples, SONAR remains the most robust retriever, achieving performance comparable to the 20K-sample setting.

| | 0-shot | | | | | | 5-shot | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T1 | T2 | T3 | T4 | T5 | T6 | T1 | T2 | T3 | T4 | T5 | T6 |
| BLOOM 7B1 | | | | | | | | | | | | |
| eng→fra | 2.6 | 14.1 | 10.5 | 22.8 | 27.5 | **41.7** | 46.6 | 46.9 | 46.4 | 46.6 | 46.7 | **47.0** |
| eng→deu | 2.1 | 2.3 | 3.1 | 6.4 | **6.6** | 1.3 | 14.0 | 14.0 | 13.5 | 13.8 | 13.9 | **14.1** |
| eng→swh | 1.4 | 1.7 | 1.5 | 1.6 | 3.2 | **3.9** | **10.8** | 10.7 | 10.5 | 10.4 | 10.5 | 10.2 |
| eng→wol | 1.3 | 1.3 | 1.7 | 1.7 | **2.5** | 0.5 | 1.5 | 1.5 | 1.4 | 1.4 | 1.6 | **1.8** |
| Mistral 7B v0.1 | | | | | | | | | | | | |
| eng→fra | 8.9 | 14.4 | 26.4 | 24.2 | **44.6** | 40.8 | **48.3** | 48.1 | 47.0 | 46.8 | 48.0 | 48.1 |
| eng→deu | 7.8 | 12.2 | 14.6 | 16.5 | **33.0** | 31.7 | 37.4 | **37.6** | 35.2 | 35.2 | 37.3 | 37.3 |
| eng→swh | **2.8** | 2.7 | 1.3 | 1.5 | 2.4 | 2.7 | 2.7 | 2.8 | 2.8 | **2.9** | 2.8 | 2.8 |
| eng→wol | **2.8** | 2.8 | 0.2 | 0.2 | 2.6 | 0.7 | 2.2 | 2.2 | 1.8 | 1.7 | **2.3** | 2.1 |
| LLaMA 2 7B | | | | | | | | | | | | |
| eng→fra | 10.2 | 11.0 | 19.3 | **28.2** | 5.3 | 8.4 | **45.4** | 45.3 | 41.3 | 41.3 | 45.2 | 45.3 |
| eng→deu | 9.8 | 12.5 | 15.1 | **19.4** | 5.1 | 3.8 | 35.2 | 35.2 | 30.0 | 31.1 | **35.2** | 34.9 |
| eng→swh | 1.1 | 1.3 | 1.0 | 0.9 | **1.3** | 0.9 | 2.7 | **2.8** | 1.6 | 0.7 | **2.8** | 2.7 |
| eng→wol | **1.5** | 1.5 | 0.0 | 0.0 | 0.2 | 0.2 | 2.1 | 2.1 | 1.5 | 1.5 | 2.1 | **2.2** |

Table 6: Comparison of BLEU scores on the FLORES-200 devtest set with three LLMs and the six templates (T1–T6) detailed in Table 5 for 0-shot and 5-shot settings. 5-shot examples are sampled uniformly at random. We report the average BLEU score across three runs with different seeds.

### C.2 Impact of in-context example order

We investigated how the ranking of in-context examples impacts translation performance. Given the huge number of permutations possible, we could not evaluate each of them. Instead, we compared the current order to its direct opposite (i.e. ranking the retrieved in-context examples from the least to the most similar starting from the source sentence). The results, given in Table 12 show that there is no significant difference in performance between the two orders.

### C.3 Impact of the domain of the selection pool

We have mainly used FLORES 200 (news domain; Goyal et al., 2022; Costa-jussà et al., 2022) for our experiments. In this section, we explore a different domain, the health domain using the TICO-19 benchmark (Anastasopoulos et al., 2020). It is a benchmark of sentences related to the COVID-19 pandemic, covering covers 35 languages. It is divided into 2 sets: an evaluation set of 2100 sentences and a validation set of 971 sentences. We evaluate on the evaluation set and study two scenarios. The in-domain scenario where the selection pool is the TICO-19's validation set which shares the same domain as the evaluation set and the out-of-domain scenario where we use FLORES 200 dev set (news domain) as the selection pool. We evaluate LLaMA 3 8B (Dubey et al., 2024) in the 5-shot setting on the translation from English to Hausa, Nepali, Somali and Urdu. In Table 13, we observe that in-domain scores are greater than their out-of-domain counteparts. However, in both scenarios, example selection via similarity search outperforms random selection. We also note that the performance gap with random selection is greater in the in-domain scenario than in the out-of-domain scenario.

### C.4 Overlap between sentence embeddings

Motivated by the low variability in performance observed between the sentence embeddings in Table 2, we analyzed the degree of overlap in the choices made by the different sentence embedding methods by calculating the average intersection between the top 10 pairs retrieved (in $\mathcal{P}$) between methods (the pool being the Flores-200 devtest set). The results in Figure 5 show that each method retrieved a distinct set of examples, with most overlap seen between E5 and Embed v3 with an average of 5.87 examples in common per top 10.

### C.5 BLEU and COMET results

As mentioned previously we additionally present results with BLEU (Table 14 and Table 16) and COMET (Table 15 and Table 17) for transparency reasons. The results show the same pattern as the laCOMET results shown in the main part of the paper. Example retrieval with sentence embeddings outperforms random sampling in all scenarios.

| | 0-shot | | | | | | 5-shot | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T1 | T2 | T3 | T4 | T5 | T6 | T1 | T2 | T3 | T4 | T5 | T6 |
| BLOOM 7B1 | | | | | | | | | | | | |
| eng→fra | 2.1 | 23.0 | 19.8 | 43.9 | 54.7 | **82.4** | 86.5 | 86.5 | 86.5 | 86.5 | 86.5 | **86.6** |
| eng→deu | 0.9 | 1.9 | 7.4 | 21.6 | **26.9** | 5.3 | **58.4** | **58.4** | 56.5 | 57.7 | 57.7 | 57.9 |
| eng→swh | 0.0 | 2.3 | 1.1 | 2.8 | 5.9 | **21.3** | 66.6 | 66.6 | 63.4 | 63.9 | 65.1 | **66.8** |
| eng→wol | 0.0 | 0.0 | 0.3 | 0.3 | **0.9** | 0.0 | **46.0** | **46.0** | 43.2 | 43.2 | 45.1 | 42.3 |
| Mistral 7B v0.1 | | | | | | | | | | | | |
| eng→fra | 11.9 | 21.5 | 61.2 | 62.0 | **81.8** | 75.4 | **86.6** | 86.5 | 85.7 | 85.6 | 86.5 | **86.6** |
| eng→deu | 13.6 | 24.8 | 47.0 | 55.6 | **78.3** | 74.4 | 85.4 | 85.4 | 83.8 | 83.5 | 85.4 | 85.2 |
| eng→swh | 1.9 | 2.3 | 13.0 | 15.5 | 26.8 | **29.3** | 33.7 | **33.8** | 32.1 | 32.8 | 33.5 | 33.3 |
| eng→wol | 0.0 | 0.0 | 1.1 | 1.1 | **1.7** | 0.0 | 34.3 | 34.3 | 38.0 | **38.1** | 33.1 | 33.7 |
| LLaMA 2 7B | | | | | | | | | | | | |
| eng→fra | 18.7 | 18.4 | 49.5 | **66.6** | 5.7 | 17.6 | 85.9 | 85.9 | 81.1 | 81.7 | 85.9 | **86.0** |
| eng→deu | 23.5 | 29.1 | 42.5 | **57.7** | 7.3 | 4.2 | 84.6 | 84.8 | 75.0 | 77.8 | 84.8 | **85.0** |
| eng→swh | 1.2 | 0.4 | 3.8 | **5.3** | 0.2 | 0.5 | 34.2 | 34.3 | 9.8 | 6.2 | 34.6 | **35.5** |
| eng→wol | 0.0 | 0.0 | **0.1** | **0.1** | 0.0 | 0.0 | 34.4 | 34.4 | 31.4 | 31.4 | 34.7 | **37.8** |

Table 7: Comparison of laCOMET scores on the FLORES-200 devtest set with three LLMs and the six templates (T1–T6) detailed in Table 5 for 0-shot and 5-shot settings. 5-shot examples are sampled uniformly at random. We report the average BLEU score across three runs with different seeds.

| | |
|---|---|
| Source sentence | International sanctions have meant that new aircraft cannot be purchased. |
| 0-shot translation (paraphrases source) | Senegal is under international sanctions, so new aircraft cannot be purchased. |
| Source sentence | During his trip, Iwasaki ran into trouble on many occasions. |
| 0-shot translation (wrong language) | Durant son voyage, Iwasaki a rencontré beaucoup de problèmes. |

Table 8: Examples of 0-shot eng→wol mistranslations by Mixtral 8x7B v0.1 with Template T5.

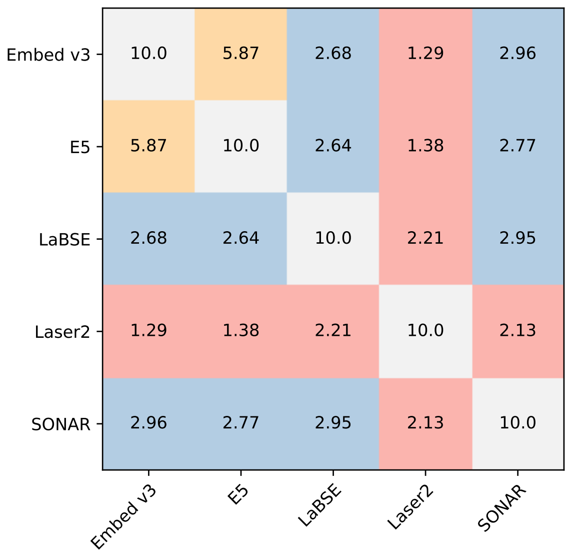


Figure 5: Average number of retrieved examples in common between sentence embedding methods (10-shot).

## C.6 Additional results for other LLMs

In Tables 18 and 19, we provide the laCOMET scores for six additional LLMs: Gemma 2B, Gemma 7B, OLMo 7B, LLaMA 2 13B, LLaMA 2 70B, and Mixtral 8x7B v0.1. We observe the same results as with BLOOM 7B1, Mistral 7B v0.1, LLaMA 2 7B and Gemma 7B. Example retrieval with sentence embeddings outperforms random sampling at all scales, with the delta being higher when translating into Swahili and Wolof. SONAR is overall the best alternative, followed by BM25.

## C.7 Additional results for more target languages

In this section, we run additional experiments to compare example retrieval via similarity search with BM25 and SONAR to random sampling in the 5-shot scenario. We use LLaMA 3 8B (Dubey et al., 2024) and we consider 8 more FLORES-200 languages: Hausa (hau), Javanese (jav), Somali (som), Telugu (tel), Urdu (urd), Xhosa (xho), Yoruba (yor) and Zulu (zul). We report the BLEU and laCOMET scores in Table 20. We observe the same results as those documented throughout the paper, choosing in-context demonstrations based on their similarity to the sentence to translate yields better translations as evaluated by BLEU and laCOMET. In table 21 we observe than even when we change the source

| | 0-shot | | | | | | 5-shot | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T1 | T2 | T3 | T4 | T5 | T6 | T1 | T2 | T3 | T4 | T5 | T6 |
| BLOOM 7B1 | | | | | | | | | | | | |
| eng→fra | 57.4 | 68.3 | 65.4 | 74.6 | 76.0 | **83.5** | 86.5 | 86.5 | 86.5 | 86.5 | 86.5 | **86.6** |
| eng→deu | 52.8 | 54.1 | 53.8 | **57.8** | 55.9 | 42.6 | **58.8** | **58.8** | 58.0 | 58.4 | 58.5 | 58.5 |
| eng→swh | 44.5 | 46.8 | 47.3 | 47.4 | 52.8 | 54.5 | **67.9** | **67.9** | 67.5 | 67.4 | 67.7 | **67.9** |
| eng→wol | 28.6 | 28.6 | 29.6 | 29.6 | 29.5 | **35.8** | **47.5** | **47.5** | 45.4 | 45.4 | 47.3 | 46.9 |
| Mistral 7B v0.1 | | | | | | | | | | | | |
| eng→fra | 66.4 | 69.4 | 67.1 | 65.6 | **84.7** | 83.4 | **86.7** | **86.7** | 86.0 | 85.8 | **86.7** | **86.7** |
| eng→deu | 61.6 | 66.2 | 54.3 | 60.9 | **82.4** | 81.7 | 85.6 | 85.6 | 84.3 | 84.1 | **85.7** | 85.5 |
| eng→swh | 49.8 | **50.3** | 36.8 | 36.6 | 36.5 | 39.1 | 36.5 | 36.5 | 36.1 | 36.3 | **36.6** | 36.1 |
| eng→wol | 29.3 | 29.3 | 22.0 | 22.0 | 29.2 | **34.5** | 40.7 | 40.7 | 41.5 | **41.6** | 40.3 | 40.5 |
| LLaMA 2 7B | | | | | | | | | | | | |
| eng→fra | 59.9 | 63.5 | 61.5 | **71.1** | 56.9 | 58.3 | 85.9 | 85.9 | 82.0 | 82.2 | 85.9 | **86.0** |
| eng→deu | 54.6 | 61.3 | 58.1 | **65.3** | 52.3 | 53.9 | 84.8 | 84.9 | 79.7 | 80.7 | 84.8 | **85.1** |
| eng→swh | 38.1 | 37.2 | 31.7 | 32.2 | **40.6** | 36.0 | 37.4 | 37.3 | **40.3** | 39.6 | 37.5 | 37.6 |
| eng→wol | 25.7 | 25.7 | 29.1 | 29.1 | 28.0 | **37.1** | 41.8 | 41.8 | 38.5 | 38.5 | 42.2 | **44.3** |

Table 9: Comparison of COMET scores on the FLORES-200 devtest set with three LLMs and the six templates (T1–T6) detailed in Table 5 for 0-shot and 5-shot settings. 5-shot examples are sampled uniformly at random. We report the average BLEU score across three runs with different seeds.

| | 0-shot | | | | | | 5-shot | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T1 | T2 | T3 | T4 | T5 | T6 | T1 | T2 | T3 | T4 | T5 | T6 |
| BLOOM 7B1 | | | | | | | | | | | | |
| eng→swh | 14.49 | 13.37 | 12.98 | 13.24 | **9.77** | 11.16 | 8.71 | 8.55 | **8.46** | 8.52 | 8.55 | 9.09 |
| eng→wol | 23.76 | 23.76 | 23.81 | 23.81 | **23.29** | 24.4 | 24.86 | 24.86 | 24.83 | 24.83 | 24.79 | **24.69** |

Table 10: Comparison of Metric-X23 (Juraska et al., 2023) scores on the FLORES-200 devtest set with BLOOM 7B1 and the six templates (T1–T6) detailed in Table 5 for 0-shot and 5-shot settings. 5-shot examples are sampled uniformly at random. We report the average MetricX-23 Large score across three runs with different seeds. Lower is better.

language from English to French or German, example selection via similarity search yields considerably better results compared to random sampling (in terms of BLEU).

| Target | Hau | Jav | Som | Tel | Urd | Xho | Yor | Zul |
|---|---|---|---|---|---|---|---|---|
| laCOMET | | | | | | | | |
| Random | 63.0 | 77.9 | 46.1 | 59.0 | 72.3 | 19.8 | 50.8 | 40.2 |
| BM25 | **65.8** | **80.5** | 51.6 | 59.5 | 73.2 | 29.3 | 53.7 | **45.8** |
| SONAR | **65.8** | **80.5** | **53.4** | **60.2** | **73.5** | **30.9** | **54.8** | 45.3 |
| BLEU | | | | | | | | |
| Random | 10.1 | 16.8 | 3.2 | 7.6 | 16.8 | 2.7 | 3.6 | 2.4 |
| BM25 | **12.4** | **18.8** | **4.6** | 7.9 | **17.3** | 3.5 | 4.8 | 4.3 |
| SONAR | 11.8 | 18.5 | 4.5 | **8.1** | 17.2 | **3.8** | **5.0** | **4.4** |

Table 20: Full quantitative results in 8 English → X translation directions from FLORES 200 with LLaMA 3 8B in 5-shot. **Bold** entries: denote statistically significant differences with $p < 0.05$ in the paired t-test compared to Random sampling.

| Target | Hau | Jav | Som | Tel | Urd | Xho | Yor | Zul |
|---|---|---|---|---|---|---|---|---|
| French | | | | | | | | |
| SONAR | **8.92** | **13.14** | **3.67** | **17.47** | **18.44** | **13.53** | **2.74** | **3.44** |
| BM25 | 5.32 | 11.00 | 1.66 | 14.55 | 18.22 | 13.35 | 1.12 | 1.44 |
| Random | 3.07 | 8.48 | 0.68 | 12.96 | 17.58 | 11.21 | 0.50 | 0.47 |
| German | | | | | | | | |
| SONAR | **7.95** | **12.21** | **1.65** | **14.37** | **18.07** | **13.95** | **2.57** | **1.48** |
| BM25 | 3.22 | 7.88 | 0.99 | 12.63 | 18.03 | 12.70 | 0.70 | 0.74 |
| Random | 2.74 | 7.07 | 0.61 | 11.47 | 17.15 | 11.36 | 0.47 | 0.42 |

Table 21: Full BLEU results in 8 French/German → X translation directions from FLORES 200 with LLaMA 3 8B in 5-shot.

### C.8 Source-to-target example retrieval

As mentioned in the main text of the article, we mainly explored source-to-source retrieval (comparing the source sentence to the source side of pool examples). In this section, we provide results for source-to-target retrieval. Tables 22 and 23 summarize the laCOMET scores obtained using different sentence embeddings with nine LLMs. Example retrieval via similarity search outperforms

| Pool Composition | 997 + 20K | | | 997 + 100K | | | 997 + 500K | | | 997 + 1M | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| BLOOM 7B1 | | | | | | | | | | | | |
| SONAR | **54.3** | **67.0** | **68.4** | **52.1** | **64.3** | **67.3** | **52.7** | **63.8** | **66.8** | **53.3** | **64.2** | **67.1** |
| BM25 | 51.0 | 62.4 | 65.2 | 49.5 | 60.1 | 63.9 | 48.4 | 61.5 | 63.9 | 49.3 | 61.3 | 62.9 |
| Random | 47.6 | 59.1 | 61.7 | 46.0 | 57.1 | 60.1 | 45.1 | 58.1 | 60.4 | 45.1 | 58.5 | 60.3 |

Table 11: Impact of adding NLLB samples to the initial selection pool (FLORES 200 dev set) in $k$-shot settings ($k \in \{1, 5, 10\}$). We report the laCOMET scores for eng→ and BLOOM 7B1.

| | | eng→fra | | | eng→deu | | | eng→swh | | | eng→wol | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| BLEU | Original | **42.9** | **47.5** | **48.0** | **12.6** | **14.9** | **15.2** | **8.6** | **12.0** | 12.7 | **2.2** | **2.9** | **3.0** |
| | Reverse | **42.9** | 47.0 | 47.8 | **12.6** | **14.9** | **15.2** | **8.6** | 11.7 | **13.1** | **2.2** | 2.8 | **3.0** |
| COMET | Original | **84.9** | **86.8** | **86.6** | **58.9** | **60.7** | **61.3** | **64.5** | **69.5** | 70.4 | **52.0** | 51.6 | **52.5** |
| | Reverse | **84.9** | 86.6 | **86.6** | **58.9** | 60.2 | **61.3** | **64.5** | 69.2 | **70.5** | **52.0** | **51.7** | **52.5** |
| laCOMET | Original | **79.8** | **86.8** | **86.6** | **55.3** | **60.1** | 60.8 | **57.4** | **68.3** | 69.6 | **50.2** | **50.4** | **51.6** |
| | Reverse | **79.8** | 86.6 | **86.6** | **55.3** | 59.6 | **60.9** | **57.4** | 67.8 | **69.7** | **50.2** | 50.3 | **51.6** |

Table 12: Impact of the ordering of in-context examples (Original: most to least similar, Reverse: least to most similar) in $k$-shot settings ($k \in \{1, 5, 10\}$) on translation quality (BLEU, COMET and laCOMET) with BLOOM 7B1 as the translator and SONAR as the example retriever.

random sampling, with most gains observed when translating into Swahili and Wolof. SONAR does even better in this setup and we attribute this to its cross-lingual training, which covers all the languages we experiment with. Comparing example retrieval in source-to-source and source-to-target does not allow us to draw systematic conclusions. However, the performance of both approaches are similar when translating into high-resource languages. When translating into low-resource languages, some sentence embeddings tend (e.g. LaBSE) to perform worse for source-to-target than for source-to-source, which is typically related to the amount of data in the language seen during training.

### C.9 Translation into English

In this section, we benchmark example retrieval with different sentence embeddings for fra→eng, deu→eng, swh→eng and wol→eng. Tables 24, 25 and 26 respectively contain the BLEU, COMET and laCOMET scores obtained with BLOOM 7B1 and LLaMA 2 7B. In this scenario, example retrieval via similarity search also proves beneficial, especially when the source language is a mid- or low-resource language. The gains are significant, but not as high as for the opposite translation direction. In summary, the conclusions are generally consistent with those for the opposite direction.

### C.10 Distribution of issues in zero-shot and few-shot MT

A major issue when translating with LLMs is the generation of empty translations and translations in the incorrect target language (a problem that appears to decrease as the number of in-context demonstrations increases). We use Mixtral 8x7B v0.1 to translate from English into French, Swahili, and Wolof. As shown in Figure 6, when translating into French, even a single in-context demonstration ensures that the language model generates a non-empty French sentence in all cases, regardless of whether the demonstrations are chosen randomly. However, for translations into Swahili and Wolof, adding in-context examples does not entirely solve the problem of translating in an incorrect language, although the more in-context demonstrations provided, the less the problem occurs. Moreover, using SONAR and BM25 sampling methods reduces the frequency of these problems compared to random sampling. As reported in Table 27, empty translations and translations in an incorrect language occur significantly less when translating into high-resource languages. When they occur, it is in setups where the number of ICL examples is low (typically equal to one) and they completely disappear for bigger values. However, these issues are more persistent with low-resource languages. Using more ICL examples

| | Hausa | | Nepali | | Somali | | Urdu | |
|---|---|---|---|---|---|---|---|---|
| | BLEU | COMET | BLEU | COMET | BLEU | COMET | BLEU | COMET |
| **IN-DOMAIN EVALUATION** | | | | | | | | |
| SONAR | 8.60 | 65.31 | 22.94 | **76.30** | 1.42 | **49.84** | **20.38** | **73.74** |
| BM25 | **9.14** | **65.85** | **23.15** | 75.90 | 1.62 | 49.78 | 20.22 | 73.70 |
| Random | 6.38 | 60.47 | 17.27 | 72.67 | 0.89 | 41.83 | 16.17 | 71.00 |
| **OUT-OF-DOMAIN EVALUATION** | | | | | | | | |
| SONAR | **6.77** | **62.21** | **17.34** | 72.73 | 1.26 | 43.89 | **15.92** | **71.57** |
| BM25 | 6.62 | 61.88 | 17.28 | **73.23** | 1.35 | **44.70** | 15.91 | 71.26 |
| Random | 5.78 | 58.92 | 15.46 | 71.38 | 0.98 | 40.73 | 15.03 | 70.60 |

Table 13: In-domain and out-of-domain evaluation on TICO-19 (Anastasopoulos et al., 2020) with LLaMA 3 8B. We report the BLEU and COMET scores.

helps mitigate their impact, but the quality of the retriever also plays a significant role.

| | eng → fra | | | eng → deu | | | eng → swh | | | eng → wol | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| BLOOM 7B1 | | | | | | | | | | | | |
| Embed v3 | 42.3 | 47.0 | 47.5 | 12.4 | 14.7 | 15.1 | **8.9** | **12.3** | **12.7** | 1.8 | 2.3 | 2.5 |
| E5 | 42.7 | 47.2 | 47.9 | 12.5 | **14.9** | **15.3** | 8.6 | 12.1 | 12.5 | 1.9 | 2.4 | 2.6 |
| LaBSE | 42.5 | 47.3 | 47.8 | 12.6 | **14.9** | 15.2 | 8.7 | 11.7 | 12.3 | **2.4** | 2.6 | 2.9 |
| LASER2 | 42.1 | 47.4 | 47.9 | **12.8** | 14.6 | 15.0 | 8.6 | 11.6 | 11.7 | **2.4** | 2.8 | 2.9 |
| SONAR | **42.9** | **47.5** | **48.0** | 12.6 | **14.9** | 15.2 | 8.6 | 12.0 | **12.7** | 2.2 | **2.9** | **3.0** |
| Random | 40.8 | 46.7 | 47.2 | 12.3 | 13.9 | 14.0 | 8.2 | 10.5 | 11.0 | 0.9 | 1.6 | 1.9 |
| Mistral 7B v0.1 | | | | | | | | | | | | |
| Embed v3 | 47.3 | 48.4 | 48.8 | 36.4 | 38.0 | **38.6** | **3.6** | **4.9** | 5.4 | 2.8 | 3.3 | 3.7 |
| E5 | 46.9 | 48.5 | 48.7 | 36.4 | 37.9 | 38.2 | 3.5 | 4.7 | **5.5** | 2.8 | 3.3 | 3.3 |
| LaBSE | 47.4 | 48.8 | 49.0 | 36.5 | 37.8 | 37.9 | 3.3 | 4.6 | 5.1 | 3.2 | 3.3 | **3.8** |
| LASER2 | **47.5** | 48.8 | 49.0 | 36.3 | 37.4 | 37.7 | 3.1 | 4.1 | 4.7 | **3.3** | 3.3 | 3.6 |
| SONAR | 47.4 | **49.0** | **49.2** | **36.6** | **38.1** | 38.2 | 3.5 | 4.6 | 5.4 | 3.2 | **3.4** | 3.7 |
| Random | 47.2 | 48.0 | 48.4 | 36.1 | 37.3 | 37.5 | 2.8 | 2.8 | 2.9 | 2.4 | 2.3 | 2.7 |
| LLaMA 2 7B | | | | | | | | | | | | |
| Embed v3 | 44.5 | 45.8 | 46.1 | 34.7 | 35.3 | 35.4 | 2.9 | 4.1 | 4.4 | 2.0 | 3.2 | 3.4 |
| E5 | 44.8 | **46.0** | **46.3** | **34.8** | **35.9** | **35.7** | 3.1 | 3.8 | 4.4 | 2.0 | 3.0 | 3.1 |
| LaBSE | 44.4 | 45.3 | 46.0 | **34.8** | 35.6 | 35.4 | **3.2** | **4.2** | 4.3 | 2.5 | **3.6** | **3.7** |
| LASER2 | 44.8 | 45.6 | 46.1 | 34.6 | 35.7 | **35.7** | 3.1 | 3.6 | 4.0 | **2.6** | **3.6** | 3.6 |
| SONAR | **44.9** | 45.5 | 46.0 | 34.5 | 35.7 | **35.7** | 3.1 | **4.2** | **4.6** | 2.1 | 3.4 | 3.7 |
| Random | 44.6 | 45.2 | 45.4 | 34.1 | 35.2 | 35.5 | 2.4 | 2.8 | 2.8 | 1.3 | 2.1 | 2.3 |

Table 14: BLEU scores for $k$-shot ($k \in \{1, 5, 10\}$) example retrieval with different sentence embeddings.

| | eng → fra | | | eng → deu | | | eng → swh | | | eng → wol | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| BLOOM 7B1 | | | | | | | | | | | | |
| Embed v3 | 84.6 | 86.7 | **86.7** | **59.0** | 60.6 | **61.3** | **65.1** | **69.7** | 70.2 | **52.4** | 51.4 | 52.0 |
| E5 | **85.0** | 86.6 | **86.7** | 58.7 | 60.5 | 61.0 | **65.1** | 69.6 | 70.2 | 52.7 | **52.3** | 51.9 |
| LaBSE | 84.7 | 86.7 | **86.7** | 58.8 | 60.4 | 61.2 | 64.5 | 69.2 | 69.9 | 52.0 | **52.3** | **53.1** |
| LASER2 | 84.8 | 86.6 | **86.7** | 58.8 | 60.3 | 60.3 | 64.1 | 68.9 | 68.9 | 51.5 | 51.4 | 52.3 |
| SONAR | 84.9 | **86.8** | 86.6 | 58.9 | **60.7** | **61.3** | 64.5 | 69.5 | **70.4** | 52.0 | 51.6 | 52.5 |
| Random | 84.3 | 86.5 | 86.6 | 58.0 | 58.5 | 58.7 | 64.0 | 67.7 | 67.9 | 49.0 | 47.3 | 48.3 |
| Mistral 7B v0.1 | | | | | | | | | | | | |
| Embed v3 | **86.6** | **87.0** | 87.0 | 84.8 | 85.8 | 86.0 | **41.8** | 43.0 | **45.1** | 45.2 | 48.2 | 48.6 |
| E5 | 86.4 | **87.0** | 86.9 | 84.9 | 85.7 | 85.8 | 41.6 | **43.3** | 44.9 | **45.5** | **48.5** | 48.5 |
| LaBSE | 86.5 | 86.9 | 87.0 | 84.9 | 85.7 | 85.9 | 41.3 | 42.2 | 43.7 | 45.5 | 47.1 | 48.8 |
| LASER2 | 86.5 | **87.0** | 87.0 | **85.0** | 85.8 | 85.8 | 39.7 | 40.1 | 41.9 | 43.3 | 47.0 | 47.6 |
| SONAR | 86.3 | **87.0** | **87.1** | **85.0** | **85.9** | **86.1** | 41.4 | 42.8 | 45.1 | 45.3 | 48.4 | **49.0** |
| Random | 86.4 | 86.7 | 86.7 | 84.7 | 85.7 | 85.7 | 38.1 | 36.6 | 36.7 | 39.3 | 40.3 | 42.4 |
| LLaMA 2 7B | | | | | | | | | | | | |
| Embed v3 | 85.8 | 86.1 | 86.3 | 84.2 | 85.0 | 85.0 | **48.7** | **45.8** | **46.5** | 48.5 | 50.0 | **50.4** |
| E5 | 85.8 | **86.2** | **86.4** | **84.4** | 85.2 | 85.2 | 48.3 | 45.1 | 46.5 | **48.9** | **50.2** | 49.7 |
| LaBSE | 85.8 | 86.0 | 86.2 | 84.4 | 85.2 | 85.1 | 47.6 | 44.9 | 45.9 | 48.2 | 49.0 | 49.6 |
| LASER2 | 85.8 | **86.2** | 86.2 | 84.1 | 85.1 | 85.3 | 44.8 | 42.3 | 43.3 | 47.2 | 48.4 | 48.2 |
| SONAR | **85.9** | 86.1 | 86.3 | 84.2 | **85.3** | **85.4** | 48.5 | 44.8 | 46.4 | 47.9 | **50.2** | 50.3 |
| Random | 85.6 | 85.9 | 86.0 | 84.1 | 84.9 | 85.1 | 40.2 | 37.5 | 37.9 | 44.2 | 42.2 | 43.2 |

Table 15: COMET scores for $k$-shot ($k \in \{1, 5, 10\}$) example retrieval with different sentence embeddings.

| | eng→fra | | | eng→deu | | | eng→swh | | | eng→wol | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| BLOOM 7B1 | | | | | | | | | | | | |
| SONAR | 42.9 | 47.5 | 48.0 | **12.6** | 14.9 | 15.2 | 8.6 | **12.0** | 12.7 | **2.2** | **2.9** | **3.0** |
| BM25 | 41.1 | **47.7** | **48.1** | **12.6** | **15.1** | 15.2 | 8.8 | 11.6 | **12.9** | 1.8 | 2.3 | 2.8 |
| R-BM25 | **43.3** | 46.1 | 46.8 | 12.4 | 13.7 | 13.8 | 7.9 | 10.2 | 10.7 | 1.2 | 1.5 | 2.1 |
| BLEU | 41.5 | 47.4 | 47.6 | 12.4 | 14.8 | **15.3** | **8.9** | 11.4 | 12.2 | 1.5 | 2.5 | 2.8 |
| RoBERTa | 41.3 | 46.6 | 47.6 | 12.4 | 14.2 | 14.0 | 8.6 | 10.5 | 11.4 | 1.6 | 2.2 | 2.2 |
| Random | 40.8 | 46.7 | 47.2 | 12.3 | 13.9 | 14.0 | 8.2 | 10.5 | 11.0 | 0.9 | 1.6 | 1.9 |
| Mistral 7B v0.1 | | | | | | | | | | | | |
| SONAR | 47.4 | **49.0** | **49.2** | 36.6 | **38.1** | **38.2** | **3.5** | 4.6 | 5.4 | **3.2** | **3.4** | **3.7** |
| BM25 | 47.7 | 48.6 | 49.0 | 36.5 | 37.9 | 38.1 | 3.4 | **5.0** | **5.7** | 2.8 | 3.3 | 3.4 |
| R-BM25 | 47.5 | 47.8 | 48.3 | 36.4 | 36.9 | 36.9 | 2.6 | 2.9 | 2.9 | 2.5 | 2.6 | 2.9 |
| BLEU | **47.9** | 48.5 | 49.0 | **36.8** | 37.6 | 37.8 | **3.5** | 4.5 | 4.8 | 2.6 | 2.9 | 3.2 |
| RoBERTa | 47.6 | 48.6 | 49.0 | 36.3 | 37.5 | 37.8 | 2.9 | 3.3 | 3.8 | 2.6 | 2.7 | 2.8 |
| Random | 47.2 | 48.0 | 48.4 | 36.1 | 37.3 | 37.5 | 2.8 | 2.8 | 2.9 | 2.4 | 2.3 | 2.7 |
| LLaMA 2 7B | | | | | | | | | | | | |
| SONAR | 44.9 | 45.5 | 46.0 | 34.5 | 35.7 | 35.7 | **3.1** | **4.2** | 4.6 | **2.1** | **3.4** | **3.7** |
| BM25 | **45.0** | 45.9 | 46.1 | 34.4 | **35.8** | **36.1** | **3.1** | 4.0 | **4.7** | 1.8 | 3.0 | 3.0 |
| R-BM25 | 44.5 | 45.2 | 45.0 | 33.8 | 34.9 | 35.1 | 2.5 | 2.8 | 2.9 | 1.2 | 2.3 | 2.4 |
| BLEU | 44.8 | **46.0** | **46.4** | **34.6** | 35.6 | 35.7 | 3.0 | 3.9 | 4.3 | 1.7 | 2.7 | 3.1 |
| RoBERTa | 44.7 | 45.8 | 45.9 | **34.6** | 35.6 | 35.9 | 2.7 | 3.1 | 3.5 | 1.4 | 2.5 | 2.6 |
| Random | 44.6 | 45.2 | 45.4 | 34.1 | 35.2 | 35.5 | 2.4 | 2.8 | 2.8 | 1.3 | 2.1 | 2.3 |

Table 16: Comparison of $k$-shot ($k \in \{1, 5, 10\}$) example retrieval with SONAR to baseline methods (BLEU).

| | eng → fra | | | eng → deu | | | eng → swh | | | eng → wol | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| BLOOM 7B1 | | | | | | | | | | | | |
| SONAR | 84.9 | **86.8** | 86.6 | **58.9** | **60.7** | **61.3** | 64.5 | **69.5** | **70.4** | **52.0** | **51.6** | **52.5** |
| BM25 | 84.6 | 86.6 | 86.7 | 58.3 | 60.1 | 60.1 | 64.6 | 68.4 | 69.5 | 51.3 | 50.8 | 51.6 |
| R-BM25 | **85.2** | 86.4 | 86.5 | 58.0 | 58.3 | 59.2 | 63.2 | 67.4 | 67.8 | 46.7 | 46.4 | 47.7 |
| BLEU | 84.4 | 86.7 | 86.6 | 58.1 | 59.9 | 60.4 | 64.4 | 68.1 | 68.8 | 51.4 | 50.8 | 51.9 |
| RoBERTa | 84.5 | 86.7 | **86.8** | 58.5 | 59.8 | 59.1 | **64.8** | 67.7 | 68.5 | 51.7 | 50.7 | 50.8 |
| Random | 84.3 | 86.5 | 86.6 | 58.0 | 58.5 | 58.7 | 64.0 | 67.7 | 67.9 | 49.0 | 47.3 | 48.3 |
| Mistral 7B v0.1 | | | | | | | | | | | | |
| SONAR | 86.3 | **87.0** | **87.1** | **85.0** | **85.9** | **86.1** | **41.4** | **42.8** | **45.1** | **45.3** | **48.4** | **49.0** |
| BM25 | **86.6** | 86.8 | 86.9 | 84.8 | 85.7 | 85.9 | 40.1 | 41.1 | 43.2 | 43.6 | 45.8 | 47.6 |
| R-BM25 | 86.5 | 86.7 | 86.7 | 84.9 | 85.6 | 85.8 | 37.6 | 36.5 | 36.6 | 38.3 | 39.1 | 41.2 |
| BLEU | **86.6** | 86.9 | 86.9 | 84.9 | 85.7 | 85.9 | 39.8 | 39.9 | 41.1 | 42.8 | 44.8 | 46.7 |
| RoBERTa | 86.5 | 86.9 | 87.0 | 84.9 | 85.6 | 86.0 | 39.1 | 38.2 | 39.3 | 42.5 | 44.5 | 45.7 |
| Random | 86.4 | 86.7 | 86.7 | 84.7 | 85.7 | 85.7 | 38.1 | 36.6 | 36.7 | 39.3 | 40.3 | 42.4 |
| LLaMA 2 7B | | | | | | | | | | | | |
| SONAR | **85.9** | 86.1 | **86.3** | 84.2 | **85.3** | **85.4** | **48.5** | **44.8** | **46.4** | **47.9** | **50.2** | **50.3** |
| BM25 | 85.7 | 86.1 | 86.2 | 84.0 | 85.0 | 85.1 | 44.4 | 42.3 | 43.9 | 46.8 | 48.0 | 48.6 |
| R-BM25 | 85.6 | 86.0 | 85.8 | 84.0 | 85.1 | 85.0 | 39.2 | 37.2 | 37.3 | 40.7 | 38.7 | 40.9 |
| BLEU | 85.6 | 86.0 | 86.1 | **84.3** | 85.0 | 85.0 | 43.2 | 41.1 | 41.8 | 46.3 | 46.9 | 47.7 |
| RoBERTa | 85.7 | **86.2** | 86.0 | **84.3** | 85.1 | 85.3 | 44.2 | 40.1 | 41.3 | 47.0 | 47.0 | 47.2 |
| Random | 85.6 | 85.9 | 86.0 | 84.1 | 84.9 | 85.1 | 40.2 | 37.5 | 37.9 | 44.2 | 42.2 | 43.2 |

Table 17: Comparison of $k$-shot ($k \in \{1, 5, 10\}$) example retrieval with SONAR to baseline methods (COMET).

| | eng→fra | | | eng→deu | | | eng→swh | | | eng→wol | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| Gemma 2B | | | | | | | | | | | | |
| Embed v3 | 84.7 | **85.3** | **85.4** | 82.0 | 83.1 | 83.3 | 63.9 | **68.0** | **68.6** | **39.1** | 45.7 | 47.1 |
| E5 | 84.6 | 85.1 | **85.4** | 82.2 | 83.2 | 83.2 | **64.1** | 67.8 | 68.1 | 38.6 | **45.8** | 47.2 |
| LaBSE | **84.8** | 85.2 | **85.4** | **82.2** | **83.4** | 83.4 | 64.0 | 67.0 | 68.1 | 36.3 | 44.7 | 46.9 |
| LASER2 | 84.6 | 85.0 | 85.0 | 82.0 | 83.1 | 83.2 | 63.7 | 66.3 | 67.4 | 32.5 | 42.7 | 44.9 |
| SONAR | **84.8** | 85.2 | 85.3 | 82.0 | 83.2 | **83.5** | 63.7 | 67.3 | 68.5 | 38.2 | 44.5 | **47.4** |
| Random | 84.6 | 84.7 | 84.9 | 81.7 | 82.7 | 83.0 | 62.3 | 64.4 | 65.1 | 26.8 | 35.2 | 37.7 |
| Gemma 7B | | | | | | | | | | | | |
| Embed v3 | 87.5 | **88.0** | **88.1** | 86.7 | 87.3 | 87.5 | 79.0 | 80.7 | **81.4** | 39.0 | 45.2 | 48.0 |
| E5 | 87.4 | 87.9 | **88.1** | 86.9 | 87.4 | **87.6** | **79.4** | 80.5 | 81.2 | **39.5** | 45.0 | **48.4** |
| LaBSE | **87.7** | 87.9 | 88.0 | **87.1** | **87.6** | 87.3 | 79.1 | **80.8** | 81.1 | 37.0 | 44.4 | 47.8 |
| LASER2 | 87.5 | 87.9 | 87.9 | **87.1** | 87.3 | 87.2 | **79.4** | 80.6 | 80.5 | 36.0 | 43.9 | 47.6 |
| SONAR | 87.4 | **88.0** | **88.1** | 86.8 | **87.6** | **87.6** | 79.2 | 80.4 | 80.7 | 38.1 | **45.6** | 48.3 |
| Random | 87.5 | 87.9 | 88.0 | 86.6 | 87.2 | 87.3 | 78.4 | 79.6 | 79.8 | 30.9 | 37.4 | 40.5 |
| OLMo 7B | | | | | | | | | | | | |
| Embed v3 | **81.0** | 81.1 | 81.2 | **75.0** | 75.7 | 75.6 | 43.2 | 43.0 | 44.2 | 40.4 | **42.1** | 43.6 |
| E5 | **81.0** | **81.4** | 81.3 | 74.7 | 75.9 | 76.0 | 42.9 | 42.0 | 43.0 | **40.6** | 41.4 | 43.6 |
| LaBSE | **81.0** | **81.4** | 81.4 | 74.8 | 75.6 | 76.0 | 42.6 | 42.5 | 43.4 | 37.8 | 41.7 | 43.3 |
| LASER2 | 80.8 | 81.3 | **81.5** | 74.4 | **76.3** | **76.2** | 39.6 | 40.2 | 41.3 | 35.3 | 40.1 | 42.5 |
| SONAR | 80.8 | 81.3 | 81.3 | 74.9 | 75.9 | 76.0 | **43.4** | **43.4** | **44.4** | 39.7 | 41.6 | **44.5** |
| Random | 80.8 | 80.8 | 80.7 | 74.3 | 75.3 | 75.4 | 36.2 | 36.8 | 37.1 | 30.1 | 33.7 | 37.0 |
| LLaMA 2 13B | | | | | | | | | | | | |
| Embed v3 | **87.2** | 87.3 | **87.6** | 85.9 | 85.9 | 86.3 | 43.4 | 46.3 | **47.8** | 40.9 | 42.6 | 44.1 |
| E5 | 87.1 | 87.3 | 87.5 | **86.0** | 86.2 | 86.4 | **43.5** | 46.1 | 47.6 | **41.7** | 43.4 | 43.5 |
| LaBSE | **87.2** | **87.4** | 87.4 | 85.7 | 86.2 | **86.6** | 42.5 | 45.6 | 47.4 | 39.2 | 42.2 | 43.4 |
| LASER2 | 87.0 | 87.2 | 87.4 | 85.7 | **86.3** | 86.2 | 41.7 | 43.7 | 45.2 | 36.7 | 41.7 | 42.8 |
| SONAR | 87.2 | 87.1 | 87.4 | 85.7 | 86.0 | **86.6** | 43.0 | **46.4** | 47.7 | 39.6 | **44.9** | **44.5** |
| Random | 86.9 | 87.2 | 87.4 | 85.7 | 85.9 | 86.2 | 38.8 | 39.9 | 40.7 | 29.4 | 34.8 | 36.3 |
| LLaMA 2 70B | | | | | | | | | | | | |
| Embed v3 | 87.5 | 88.0 | 88.1 | **87.2** | 87.6 | 87.7 | 53.5 | **61.2** | 62.6 | 41.1 | 47.7 | **49.4** |
| E5 | **87.7** | 88.1 | **88.3** | **87.2** | 87.5 | **87.8** | 53.0 | 61.0 | **62.9** | **41.7** | **48.4** | 48.4 |
| LaBSE | 87.5 | **88.2** | 88.2 | 87.0 | 87.7 | 87.7 | **53.6** | 60.6 | 62.3 | 40.1 | 48.2 | 48.4 |
| LASER2 | 87.5 | 88.0 | 88.2 | **87.2** | 87.6 | 87.7 | 53.0 | 59.5 | 60.8 | 39.1 | 46.9 | 47.7 |
| SONAR | **87.7** | 88.1 | **88.3** | **87.2** | **87.8** | 87.7 | 52.6 | 60.8 | 62.6 | 41.3 | 48.1 | 48.9 |
| Random | 87.4 | 87.9 | 88.1 | 87.1 | 87.4 | 87.6 | 49.4 | 56.5 | 57.3 | 34.2 | 40.0 | 41.9 |
| Mixtral 8x7B v0.1 | | | | | | | | | | | | |
| Embed v3 | **88.2** | **88.4** | **88.5** | 87.6 | 87.9 | 88.1 | 53.3 | 56.9 | **59.8** | 34.1 | 45.2 | 47.9 |
| E5 | 88.0 | **88.4** | 88.4 | 87.5 | **88.2** | **88.3** | **53.5** | 56.5 | 59.4 | **34.3** | **45.3** | 47.5 |
| LaBSE | 88.2 | **88.4** | **88.5** | **87.8** | 88.1 | 88.1 | 53.1 | 56.8 | 58.8 | 32.9 | **45.3** | 47.7 |
| LASER2 | 88.0 | 88.3 | 88.4 | 87.5 | **88.2** | 88.0 | 51.5 | 55.5 | 57.6 | 32.4 | 44.5 | 47.2 |
| SONAR | **88.2** | **88.4** | **88.5** | 87.2 | 88.0 | **88.3** | 53.3 | **57.0** | 58.7 | 33.3 | 45.1 | **48.1** |
| Random | 88.0 | 88.2 | 88.3 | 87.4 | 88.0 | 88.1 | 50.3 | 52.2 | 53.5 | 25.6 | 37.9 | 40.9 |

Table 18: Additional results (other LLMs): laCOMET results for example retrieval with different sentence embeddings in $k$-shot settings ($k \in \{1, 5, 10\}$).

| | eng→fra | | | eng→deu | | | eng→swh | | | eng→wol | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| Gemma 2B | | | | | | | | | | | | |
| SONAR | **84.8** | **85.2** | **85.3** | **82.0** | 83.2 | **83.5** | 63.7 | **67.3** | **68.5** | **38.2** | **44.5** | **47.4** |
| BM25 | 84.7 | 85.1 | 85.2 | 81.9 | 83.0 | 83.1 | **64.1** | **67.3** | 68.4 | 36.3 | 43.4 | 45.3 |
| R-BM25 | 84.5 | 84.9 | 84.8 | **82.0** | 83.0 | 83.1 | 63.1 | 64.5 | 65.1 | 24.4 | 33.3 | 35.9 |
| BLEU | 84.7 | 85.0 | 85.1 | 81.8 | 83.2 | 82.9 | 63.6 | 67.0 | 67.0 | 34.1 | 42.1 | 43.3 |
| RoBERTa | **84.8** | 85.0 | 85.0 | 81.8 | **83.3** | **83.5** | 63.3 | 65.8 | 66.0 | 31.9 | 40.3 | 43.3 |
| Random | 84.6 | 84.7 | 84.9 | 81.7 | 82.7 | 83.0 | 62.3 | 64.4 | 65.1 | 26.8 | 35.2 | 37.7 |
| Gemma 7B | | | | | | | | | | | | |
| SONAR | 87.4 | 88.0 | **88.1** | 86.8 | **87.6** | **87.6** | **79.2** | **80.4** | 80.7 | **38.1** | **45.6** | **48.3** |
| BM25 | 87.6 | 88.0 | 87.7 | 86.8 | 87.2 | 87.0 | **79.2** | 80.3 | **80.9** | 35.8 | 43.6 | 47.1 |
| R-BM25 | 87.6 | 87.9 | 87.7 | 86.8 | 87.1 | 86.8 | 78.3 | 79.7 | 79.6 | 28.2 | 36.2 | 39.1 |
| BLEU | **87.7** | 87.9 | **88.1** | **87.0** | 87.4 | 87.4 | 78.9 | **80.4** | 80.2 | 34.7 | 42.0 | 45.5 |
| RoBERTa | 87.4 | **88.1** | **88.1** | 86.7 | 87.3 | 87.4 | 78.8 | 80.2 | 80.1 | 35.6 | 40.6 | 44.0 |
| Random | 87.5 | 87.9 | 88.0 | 86.6 | 87.2 | 87.3 | 78.4 | 79.6 | 79.8 | 30.9 | 37.4 | 40.5 |
| OLMo 7B | | | | | | | | | | | | |
| SONAR | 80.8 | 81.3 | **81.3** | **74.9** | **75.9** | **76.0** | **43.4** | **43.4** | **44.4** | **39.7** | **41.6** | **44.5** |
| BM25 | 80.7 | **81.4** | 81.1 | 74.6 | 75.5 | 75.7 | 40.4 | 41.1 | 42.1 | 36.9 | 40.3 | 42.3 |
| R-BM25 | 80.2 | 80.7 | 80.8 | 74.3 | 75.1 | 75.1 | 35.6 | 36.6 | 37.0 | 24.6 | 30.3 | 34.7 |
| BLEU | **80.9** | 81.1 | 81.0 | **74.9** | 75.3 | 75.8 | 39.8 | 40.5 | 41.1 | 35.5 | 40.2 | 42.4 |
| RoBERTa | 80.8 | 81.0 | 80.9 | 74.4 | 75.6 | 75.2 | 39.7 | 38.6 | 39.3 | 35.8 | 37.9 | 39.9 |
| Random | 80.8 | 80.8 | 80.7 | 74.3 | 75.3 | 75.4 | 36.2 | 36.8 | 37.1 | 30.1 | 33.7 | 37.0 |
| LLaMA 2 13B | | | | | | | | | | | | |
| SONAR | **87.2** | 87.1 | 87.4 | 85.7 | 86.0 | **86.6** | **43.0** | **46.4** | **47.7** | **39.6** | **44.9** | **44.5** |
| BM25 | 86.9 | 87.0 | 87.3 | **86.1** | 85.8 | 86.5 | 41.2 | 44.9 | 46.4 | 38.3 | 41.5 | 43.4 |
| R-BM25 | 87.1 | **87.2** | 87.3 | 85.9 | **86.1** | 86.4 | 38.5 | 38.9 | 40.1 | 27.3 | 33.2 | 34.6 |
| BLEU | 87.0 | 86.4 | 87.3 | 85.7 | 85.5 | 86.5 | 40.8 | 44.0 | 44.6 | 36.0 | 41.9 | 42.7 |
| RoBERTa | 87.0 | 87.0 | **87.5** | 85.9 | 85.7 | 86.3 | 40.6 | 42.2 | 43.2 | 36.9 | 39.7 | 40.3 |
| Random | 86.9 | **87.2** | 87.4 | 85.7 | 85.9 | 86.2 | 38.8 | 39.9 | 40.7 | 29.4 | 34.8 | 36.3 |
| LLaMA 2 70B | | | | | | | | | | | | |
| SONAR | **87.7** | **88.1** | **88.3** | **87.2** | **87.8** | **87.7** | **52.6** | **60.8** | **62.6** | **41.3** | **48.1** | **48.9** |
| BM25 | **87.7** | 87.9 | 88.1 | 86.9 | 87.6 | 87.7 | 50.6 | 60.1 | 61.8 | 37.9 | 45.8 | 48.7 |
| R-BM25 | 87.3 | 87.8 | 88.0 | 87.1 | 87.5 | 87.6 | 47.0 | 56.1 | 57.7 | 29.9 | 38.4 | 41.3 |
| BLEU | 87.2 | 88.0 | 88.1 | **87.2** | 87.4 | 87.6 | 50.7 | 59.4 | 60.1 | 38.7 | 45.8 | 46.5 |
| RoBERTa | 87.4 | 87.9 | 88.2 | 87.1 | 87.5 | 87.5 | 51.8 | 58.0 | 59.0 | 39.4 | 44.8 | 45.8 |
| Random | 87.4 | 87.9 | 88.1 | 87.1 | 87.4 | 87.6 | 49.4 | 56.5 | 57.3 | 34.2 | 40.0 | 41.9 |
| Mixtral 8x7B v0.1 | | | | | | | | | | | | |
| SONAR | **88.2** | 88.4 | **88.5** | 87.2 | **88.0** | **88.3** | **53.3** | **57.0** | **58.7** | **33.3** | **45.1** | **48.1** |
| BM25 | 87.9 | 88.3 | 88.2 | **87.6** | **88.0** | 88.1 | 52.4 | 56.9 | 58.0 | 30.6 | 44.9 | 46.8 |
| R-BM25 | 88.0 | 88.2 | 88.3 | 87.4 | 87.9 | 88.0 | 50.2 | 52.9 | 53.6 | 22.8 | 34.5 | 37.5 |
| BLEU | 87.8 | 88.4 | 88.4 | 87.4 | **88.0** | 88.1 | 51.4 | 55.9 | 56.9 | 30.6 | 43.4 | 46.3 |
| RoBERTa | 88.1 | **88.5** | **88.5** | 87.4 | 87.9 | 88.0 | 51.9 | 54.4 | 55.2 | 31.1 | 41.7 | 45.2 |
| Random | 88.0 | 88.2 | 88.3 | 87.4 | **88.0** | 88.1 | 50.3 | 52.2 | 53.5 | 25.6 | 37.9 | 40.9 |

Table 19: Comparison of $k$-shot ($k \in \{1, 5, 10\}$) example retrieval with SONAR to baseline methods (laCOMET).

| | eng→fra | | | eng→deu | | | eng→swh | | | eng→wol | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| BLOOM 7B1 | | | | | | | | | | | | |
| Embed v3 | 79.9 | 86.7 | **86.8** | 55.7 | **60.4** | 60.9 | 58.0 | 68.3 | 68.9 | 48.4 | 50.0 | 50.6 |
| E5 | **80.0** | 86.5 | 86.6 | 54.7 | 59.9 | 60.5 | 58.8 | 67.6 | 69.0 | 47.8 | 49.0 | 49.9 |
| LaBSE | 79.2 | 86.6 | 86.6 | 54.9 | 60.1 | 60.5 | 57.9 | **68.5** | **69.4** | 46.4 | 47.4 | 48.6 |
| LASER2 | 78.7 | **86.9** | 86.7 | 54.6 | 60.0 | 59.9 | **58.9** | 67.7 | 68.3 | 50.4 | 50.8 | 51.0 |
| SONAR | 79.8 | 86.6 | 86.6 | **55.9** | 60.1 | **61.5** | 57.8 | 68.1 | 68.9 | **50.9** | **51.2** | **52.1** |
| Random | 77.3 | 86.5 | 86.6 | 52.8 | 57.7 | 57.7 | 56.9 | 65.1 | 66.0 | 46.5 | 45.1 | 46.4 |
| Mistral 7B v0.1 | | | | | | | | | | | | |
| Embed v3 | 85.9 | **87.0** | **87.0** | 83.2 | 85.7 | 85.7 | 37.0 | **41.3** | **44.3** | 34.4 | 41.8 | 43.2 |
| E5 | 86.0 | 86.5 | **87.0** | 82.7 | 85.4 | 85.7 | 36.8 | 40.9 | 42.0 | 34.4 | 40.8 | 43.9 |
| LaBSE | **86.2** | **87.0** | 86.9 | **83.9** | 85.3 | 85.7 | 37.2 | 39.6 | 42.8 | 28.0 | 37.6 | 40.3 |
| LASER2 | **86.2** | 86.8 | 86.9 | 83.7 | 85.7 | 85.7 | 34.7 | 37.6 | 39.0 | 32.4 | 41.9 | 42.8 |
| SONAR | 86.1 | 86.8 | **87.0** | 83.6 | **85.8** | **86.0** | **37.4** | 40.9 | 42.8 | **35.3** | **44.1** | **44.5** |
| Random | 85.8 | 86.5 | 86.6 | 83.0 | 85.4 | 85.5 | 32.7 | 33.5 | 33.8 | 26.7 | 33.2 | 36.0 |
| LLaMA 2 7B | | | | | | | | | | | | |
| Embed v3 | 85.7 | 86.2 | **86.3** | 84.0 | 85.1 | **85.3** | **46.3** | **44.2** | **45.8** | 37.9 | 43.1 | 45.1 |
| E5 | **85.8** | 86.1 | **86.3** | 83.8 | 84.8 | 85.0 | 44.8 | 43.2 | 45.0 | 37.5 | 41.7 | 44.9 |
| LaBSE | 85.5 | 86.2 | **86.3** | **84.1** | 85.0 | **85.3** | 43.7 | 42.6 | 45.2 | 33.7 | 38.5 | 39.2 |
| LASER2 | **85.8** | 86.1 | 86.1 | 83.9 | 85.2 | **85.2** | 40.6 | 38.8 | 40.9 | 41.2 | 43.8 | 45.2 |
| SONAR | 85.7 | **86.3** | **86.3** | 84.0 | 85.1 | 85.2 | 45.8 | 43.2 | 45.4 | **40.8** | **45.1** | **46.3** |
| Random | 85.6 | 85.9 | 86.0 | 83.6 | 84.8 | 85.0 | 35.4 | 34.7 | 35.8 | 34.4 | 34.7 | 36.5 |
| Gemma 7B | | | | | | | | | | | | |
| Embed v3 | **87.7** | 88.0 | 88.0 | 86.8 | 87.3 | 87.6 | **79.4** | **80.7** | 80.7 | 35.6 | 43.0 | 46.5 |
| E5 | 87.6 | 87.9 | **88.1** | 86.6 | 87.4 | 87.6 | **79.4** | 80.5 | 80.8 | 35.6 | 42.3 | 46.0 |
| LaBSE | 87.6 | **88.1** | 87.9 | 87.0 | **87.6** | 87.6 | 79.1 | 80.4 | **81.0** | 33.5 | 41.7 | 44.7 |
| LASER2 | 87.5 | 88.0 | 88.3 | **87.1** | 87.5 | **87.7** | 79.1 | 79.9 | 80.6 | 33.9 | 42.6 | 46.0 |
| SONAR | 87.6 | 88.0 | **88.1** | 86.7 | 87.5 | **87.7** | **79.4** | 80.3 | 80.7 | **37.0** | **44.1** | **47.7** |
| Random | 87.5 | 87.9 | 88.0 | 86.6 | 87.2 | 87.3 | 78.4 | 79.6 | 79.8 | 30.9 | 37.4 | 40.5 |

Table 22: laCOMET scores of $k$-shot ($k \in \{1, 5, 10\}$) *source-to-target* example retrieval with different sentence embeddings for 4 LLMs (BLOOM 7B1, Mistral 7B v0.1, LLaMA 2 7B and Gemma 7B).

| | eng → fra | | | eng → deu | | | eng → swh | | | eng → wol | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| Gemma 2B | | | | | | | | | | | | |
| Embed v3 | 84.8 | **85.1** | **85.4** | **82.2** | 83.2 | 83.2 | **64.1** | 66.7 | **68.6** | 34.3 | **43.0** | 45.4 |
| E5 | **84.9** | **85.1** | 85.3 | 81.8 | 82.8 | 83.1 | 63.8 | 66.8 | 68.2 | **34.5** | 42.3 | **45.6** |
| LaBSE | 84.7 | **85.1** | 85.2 | 82.1 | **83.4** | **83.6** | 63.7 | **67.3** | 67.8 | 29.5 | 38.9 | 41.1 |
| LASER2 | 84.7 | **85.1** | 85.2 | **82.2** | 83.2 | 83.4 | 63.2 | 66.1 | 66.5 | 33.8 | 42.6 | 44.8 |
| SONAR | 84.7 | **85.1** | 85.2 | **82.2** | 83.2 | 83.4 | 63.2 | 66.1 | 66.5 | 33.8 | 42.6 | 44.8 |
| Random | 84.6 | 84.7 | 84.9 | 81.7 | 82.7 | 83.0 | 62.3 | 64.4 | 65.1 | 26.8 | 35.2 | 37.7 |
| OLMo 7B | | | | | | | | | | | | |
| Embed v3 | **81.0** | 81.3 | **81.3** | 74.7 | 75.6 | 75.7 | 43.0 | **43.3** | **44.2** | 37.0 | 40.3 | 42.4 |
| E5 | 80.9 | **81.5** | **81.3** | 74.5 | 75.5 | 75.4 | 42.4 | 42.5 | 43.7 | 37.4 | 38.5 | 41.1 |
| LaBSE | 80.8 | 81.3 | 81.2 | 74.8 | **76.0** | **76.0** | 41.8 | 42.1 | 43.8 | 31.7 | 37.8 | 40.5 |
| LASER2 | 80.8 | 81.4 | 81.1 | **74.9** | 75.8 | 75.9 | 39.6 | 39.9 | 41.0 | 35.1 | 39.5 | 41.0 |
| SONAR | **81.0** | 81.1 | 81.0 | **74.9** | 75.7 | 75.8 | **43.8** | 42.9 | 43.9 | **38.9** | **42.2** | **43.1** |
| Random | 80.8 | 80.8 | 80.7 | 74.3 | 75.3 | 75.4 | 36.2 | 36.8 | 37.1 | 30.1 | 33.7 | 37.0 |
| LLaMA 2 13B | | | | | | | | | | | | |
| Embed v3 | **87.2** | 87.2 | **87.6** | 85.9 | 86.1 | 86.5 | 42.1 | 46.1 | 47.4 | 37.9 | 42.2 | 42.6 |
| E5 | 87.1 | 86.9 | 87.3 | 85.6 | 86.1 | 86.3 | 42.3 | 45.8 | 47.2 | 37.4 | 40.9 | 42.0 |
| LaBSE | 87.1 | **87.4** | 87.5 | **86.1** | **86.3** | **86.6** | 42.8 | 45.8 | 47.7 | 32.7 | 40.6 | 40.6 |
| LASER2 | 87.1 | 87.3 | 87.5 | 85.9 | 86.2 | 86.4 | 40.3 | 43.1 | 43.9 | 35.0 | 40.4 | 41.2 |
| SONAR | **87.2** | 87.0 | 87.5 | 85.7 | 86.0 | 86.4 | **43.0** | **46.6** | **48.4** | **39.7** | **43.9** | **44.8** |
| Random | 86.9 | 87.2 | 87.4 | 85.7 | 85.9 | 86.2 | 38.8 | 39.9 | 40.7 | 29.4 | 34.8 | 36.3 |
| LLaMA 2 70B | | | | | | | | | | | | |
| Embed v3 | 87.6 | 88.1 | 88.2 | 87.1 | 87.3 | **87.8** | 53.3 | 61.0 | 62.3 | 38.9 | 46.7 | 47.5 |
| E5 | 87.7 | 88.0 | 88.2 | 87.0 | 87.5 | 87.6 | 52.0 | 60.5 | 62.4 | 38.6 | 45.7 | 47.7 |
| LaBSE | **87.8** | **88.2** | 88.2 | 87.3 | 87.5 | 87.6 | **53.5** | 60.3 | 62.3 | 37.2 | 44.0 | 46.0 |
| LASER2 | 87.5 | **88.2** | 88.2 | **87.4** | **87.7** | **87.8** | 51.2 | 59.0 | 60.1 | 40.0 | 46.1 | 46.5 |
| SONAR | 87.7 | **88.2** | **88.3** | 87.2 | 87.5 | 87.6 | 52.5 | **61.6** | **62.9** | **41.7** | **48.2** | **49.5** |
| Random | 87.4 | 87.9 | 88.1 | 87.1 | 87.4 | 87.6 | 49.4 | 56.5 | 57.3 | 34.2 | 40.0 | 41.9 |
| Mixtral 8x7B v0.1 | | | | | | | | | | | | |
| Embed v3 | **88.3** | 88.4 | 88.4 | 87.4 | **88.1** | **88.3** | **53.4** | 57.1 | **59.4** | 31.7 | 45.1 | 47.2 |
| E5 | 88.2 | 88.4 | 88.3 | 87.3 | **88.1** | 88.1 | 52.2 | 56.3 | 59.0 | 29.6 | 43.2 | 45.6 |
| LaBSE | **88.3** | 88.4 | 88.4 | **87.7** | **88.1** | 88.1 | 53.3 | 56.1 | 58.7 | 28.3 | 41.1 | 44.5 |
| LASER2 | 87.9 | 88.4 | **88.6** | 87.6 | **88.1** | 88.1 | 51.6 | 55.1 | 56.3 | 30.6 | 43.6 | 45.4 |
| SONAR | 88.2 | **88.5** | **88.6** | 87.6 | 88.0 | 88.2 | 53.3 | **57.3** | **59.4** | **34.5** | **46.1** | **47.9** |
| Random | 88.0 | 88.2 | 88.3 | 87.4 | 88.0 | 88.1 | 50.3 | 52.2 | 53.5 | 25.6 | 37.9 | 40.9 |

Table 23: Benchmarking of example retrieval *source-to-target* with different sentence embeddings in $k$-shot ($k \in \{1, 5, 10\}$). We report the laCOMET scores.

| | fra → eng | | | deu → eng | | | swh → eng | | | wol → eng | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| BLOOM 7B1 | | | | | | | | | | | | |
| Embed v3 | 44.3 | 45.2 | 45.0 | 31.2 | 31.8 | **32.4** | 27.7 | 28.8 | **28.7** | 5.6 | 6.8 | 6.8 |
| E5 | **44.4** | **45.3** | **45.5** | **31.5** | 31.9 | **32.4** | 27.6 | 28.8 | 28.5 | 5.6 | 7.1 | 6.6 |
| LaBSE | 44.1 | **45.3** | 45.2 | 31.1 | 32.0 | 32.1 | 27.7 | **29.0** | 28.4 | 5.7 | 6.8 | 6.4 |
| LASER2 | 44.2 | 44.8 | 44.6 | 31.2 | 31.4 | 31.8 | **27.8** | 28.3 | 28.3 | 5.3 | 6.8 | 6.6 |
| SONAR | 44.3 | 45.2 | 45.1 | 31.2 | **32.3** | 32.3 | 27.4 | 28.7 | 28.6 | **6.2** | **7.2** | **7.2** |
| Random | 44.0 | 45.1 | 45.0 | 30.6 | 31.2 | 31.1 | 27.6 | 28.5 | 28.4 | 5.4 | 6.7 | 6.6 |
| LLaMA 2 7B | | | | | | | | | | | | |
| Embed v3 | 44.9 | 46.4 | 46.8 | 43.7 | 45.0 | 45.6 | 9.2 | 10.9 | 11.3 | 6.1 | 7.1 | 7.2 |
| E5 | 45.1 | 46.5 | 47.0 | 43.8 | 45.5 | **45.7** | **9.4** | 11.0 | 11.3 | 6.4 | 7.3 | 7.4 |
| LaBSE | 45.3 | 46.7 | 47.2 | **43.9** | 45.0 | 45.5 | 9.2 | **11.2** | **11.4** | 6.3 | 7.2 | 7.4 |
| LASER2 | 45.0 | **46.9** | 47.1 | 43.7 | 45.3 | 45.4 | 8.7 | 10.2 | 10.5 | **6.7** | 7.4 | **7.6** |
| SONAR | **45.4** | 46.8 | **47.3** | 43.4 | **45.5** | 45.6 | 9.2 | 10.9 | **11.4** | **6.7** | **7.5** | 7.4 |
| Random | 44.5 | 45.9 | 46.6 | 43.6 | 45.1 | 45.2 | 8.7 | 9.7 | 9.8 | 6.0 | 7.0 | 6.9 |

Table 24: Benchmarking of example retrieval with different sentence embeddings in $k$-shot ($k \in \{1, 5, 10\}$). We report the BLEU scores.

| | fra → eng | | | deu → eng | | | swh → eng | | | wol → eng | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| BLOOM 7B1 | | | | | | | | | | | | |
| Embed v3 | 88.2 | **88.4** | 88.4 | **82.2** | 82.9 | **83.4** | 77.5 | 78.7 | 79.2 | 48.8 | 50.8 | 51.4 |
| E5 | 88.1 | **88.4** | 88.4 | **82.2** | 82.8 | 83.3 | 77.4 | 79.0 | 79.2 | 48.5 | 50.7 | 51.1 |
| LaBSE | **88.3** | **88.4** | 88.4 | 81.6 | 82.5 | 82.7 | 77.4 | 78.7 | 78.9 | 47.0 | 49.1 | 49.3 |
| LASER2 | **88.3** | 88.3 | 88.3 | 81.6 | 82.1 | 82.3 | 77.4 | 78.4 | 78.7 | 47.3 | 49.6 | 49.9 |
| SONAR | 88.2 | 88.3 | **88.5** | 82.0 | **83.0** | 83.2 | **77.7** | **79.1** | **79.6** | **49.2** | **51.3** | **51.7** |
| Random | 88.2 | **88.4** | 88.3 | 81.1 | 81.7 | 81.7 | 77.1 | 78.1 | 78.4 | 45.1 | 47.4 | 47.9 |
| LLaMA 2 7B | | | | | | | | | | | | |
| Embed v3 | 88.6 | 88.9 | 89.0 | **88.5** | **88.8** | 88.8 | **59.8** | **63.4** | **64.2** | 48.8 | 50.4 | 51.4 |
| E5 | 88.6 | 88.9 | 89.0 | **88.5** | 88.7 | **88.9** | 59.4 | 62.9 | 63.7 | 48.7 | 50.8 | 51.6 |
| LaBSE | **88.7** | 88.9 | 89.0 | **88.5** | **88.8** | 88.8 | 59.0 | 62.7 | 63.1 | 47.2 | 49.1 | 50.0 |
| LASER2 | **88.7** | 88.9 | 89.0 | 88.4 | **88.8** | 88.8 | 57.7 | 60.3 | 61.1 | 47.6 | 49.7 | 50.3 |
| SONAR | **88.7** | **89.0** | **89.1** | **88.5** | **88.8** | 88.8 | 59.7 | 63.3 | **64.2** | **49.2** | **51.5** | **51.9** |
| Random | 88.6 | 88.8 | 88.9 | 88.4 | 88.7 | 88.7 | 56.1 | 58.0 | 58.8 | 45.2 | 47.6 | 48.2 |

Table 25: COMET scores for $k$-shot ($k \in \{1, 5, 10\}$) example retrieval with different sentence embeddings.

| | fra → eng | | | deu → eng | | | swh → eng | | | wol → eng | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| BLOOM 7B1 | | | | | | | | | | | | |
| Embed v3 | 88.2 | 88.4 | 88.4 | 82.1 | 82.9 | **83.4** | **77.3** | 78.6 | 79.0 | 48.6 | 50.6 | 51.2 |
| E5 | 88.1 | 88.4 | 88.4 | **82.2** | 82.8 | 83.3 | 77.2 | 78.8 | 79.0 | 48.3 | 50.5 | 51.0 |
| LaBSE | 88.3 | **88.4** | 88.4 | 81.6 | 82.5 | 82.7 | 76.8 | 78.6 | 78.6 | 46.4 | 48.6 | 49.1 |
| LASER2 | **88.3** | 88.3 | 88.3 | 81.5 | 82.1 | 82.3 | 76.7 | 78.1 | 78.6 | 46.8 | 49.0 | 49.5 |
| SONAR | 88.2 | 88.3 | **88.4** | 82.0 | **83.0** | 83.2 | 77.1 | **78.9** | **79.4** | **48.7** | **51.1** | **51.6** |
| Random | 88.1 | 88.4 | 88.3 | 81.1 | 81.7 | 81.7 | 76.2 | 77.8 | 78.3 | 43.9 | 46.4 | 47.2 |
| LLaMA 2 7B | | | | | | | | | | | | |
| Embed v3 | 88.6 | 88.9 | 89.0 | **88.5** | **88.8** | 88.8 | **59.3** | **63.4** | **64.1** | **48.5** | 50.2 | 51.3 |
| E5 | 88.6 | 88.9 | 89.0 | **88.5** | 88.7 | **88.9** | 59.1 | 62.8 | 63.7 | 47.9 | 50.7 | **51.6** |
| LaBSE | **88.7** | 88.9 | 89.0 | **88.5** | **88.8** | 88.8 | 58.4 | 62.6 | 63.1 | 46.6 | 48.9 | 49.8 |
| LASER2 | **88.7** | 88.9 | 89.0 | 88.4 | **88.8** | 88.8 | 57.1 | 60.1 | 61.0 | 46.9 | 49.4 | 49.9 |
| SONAR | **88.7** | **89.0** | **89.1** | **88.5** | **88.8** | 88.8 | 59.2 | 63.2 | 64.0 | 48.4 | **51.5** | 51.5 |
| Random | 88.6 | 88.8 | 88.9 | 88.4 | 88.7 | 88.7 | 55.6 | 57.8 | 58.6 | 44.1 | 46.9 | 47.5 |

Table 26: laCOMET scores for $k$-shot ($k \in \{1, 5, 10\}$) example retrieval with different sentence embeddings for into-English language directions.

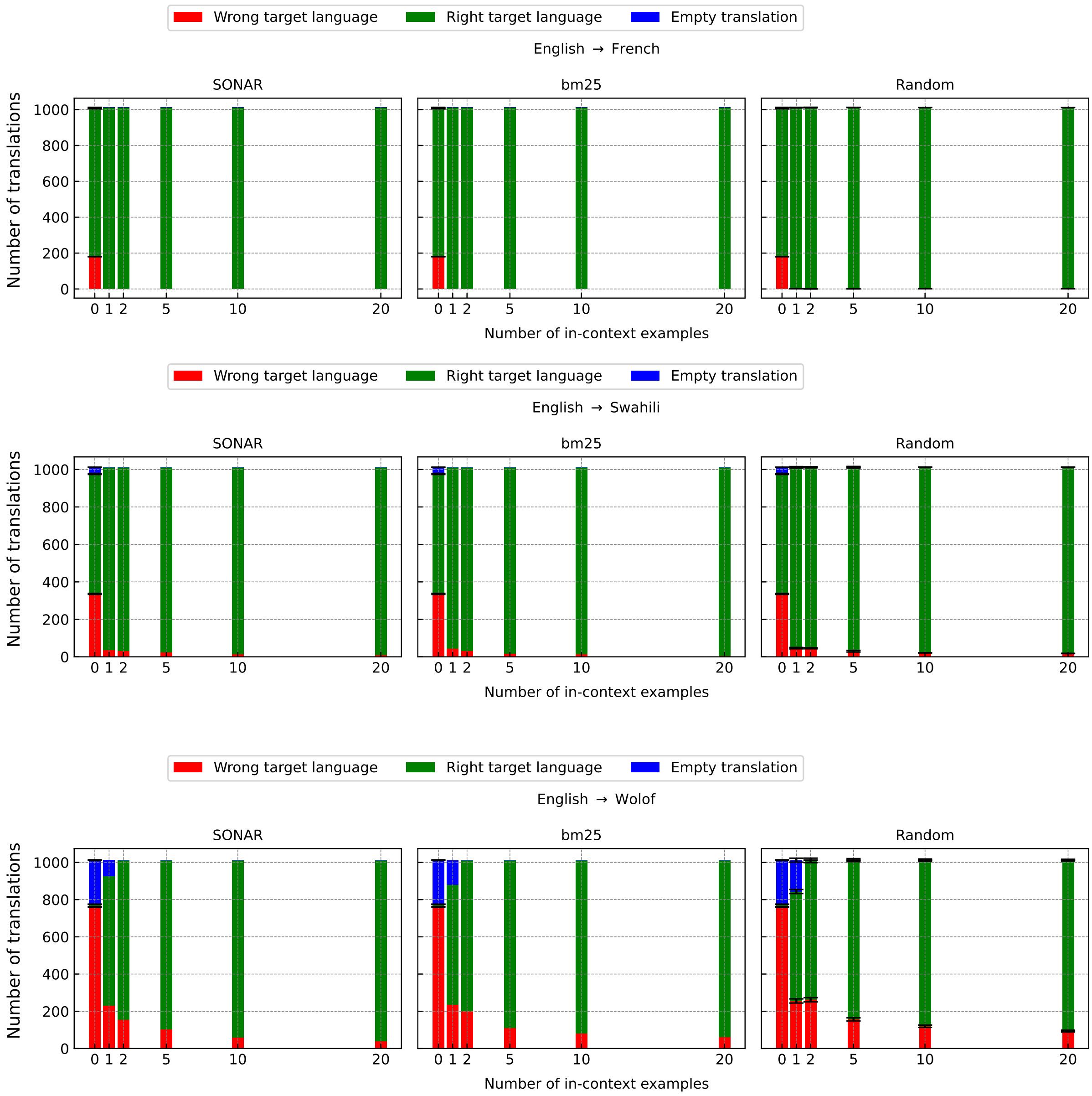


Figure 6: Error analysis of few-shot translation (eng→{fra, swa, wol}), of Mixtral 8x7B v0.1, tracking the number of empty translations, the number of translation in the wrong target language and those in the right language.

| Method | eng→fra | | | eng→deu | | | eng→swh | | | eng→wol | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 | 1 | 5 | 10 |
| BLOOM 7B1 | | | | | | | | | | | | |
| Embed v3 | 937 | 1012 | 1012 | 947 | 1001 | 1005 | 886 | 985 | 993 | 961 | 974 | 980 |
| E5 | 945 | 1012 | 1012 | 943 | 1004 | 1002 | 908 | 981 | 991 | 963 | 989 | 979 |
| LaBSE | 935 | 1012 | 1012 | 948 | 1004 | 1002 | 891 | 981 | 996 | 954 | 980 | 993 |
| LASER2 | 931 | 1012 | 1012 | 947 | 1005 | 999 | 891 | 986 | 986 | 925 | 974 | 975 |
| SONAR | 937 | 1012 | 1012 | 951 | 999 | 1002 | 874 | 988 | 994 | 960 | 976 | 990 |
| Random | 909.67 | 1012 | 1012 | 926 | 997 | 994.67 | 876 | 960 | 971.33 | 932.67 | 945 | 956.67 |
| Mistral 7B v0.1 | | | | | | | | | | | | |
| Embed v3 | 1006 | 1012 | 1011 | 990 | 1010 | 1009 | 909 | 960 | 958 | 734 | 882 | 890 |
| E5 | 1002 | 1012 | 1011 | 986 | 1004 | 1008 | 909 | 956 | 961 | 726 | 883 | 879 |
| LaBSE | 1007 | 1010 | 1012 | 986 | 1005 | 1008 | 911 | 946 | 969 | 730 | 869 | 886 |
| LASER2 | 1005 | 1011 | 1012 | 987 | 1008 | 1007 | 906 | 944 | 964 | 650 | 862 | 883 |
| SONAR | 1008 | 1010 | 1011 | 988 | 1010 | 1009 | 916 | 941 | 948 | 732 | 909 | 920 |
| Random | 1003.33 | 1008.67 | 1010.33 | 984 | 1007 | 1009 | 892.33 | 919 | 920.67 | 493.33 | 773 | 803 |
| LLaMA 2 7B | | | | | | | | | | | | |
| Embed v3 | 1012 | 1012 | 1012 | 1007 | 1011 | 1012 | 932 | 952 | 982 | 801 | 903 | 921 |
| E5 | 1012 | 1012 | 1012 | 1007 | 1012 | 1012 | 928 | 965 | 978 | 802 | 904 | 931 |
| LaBSE | 1010 | 1012 | 1012 | 1006 | 1011 | 1012 | 927 | 948 | 974 | 758 | 861 | 898 |
| LASER2 | 1012 | 1012 | 1012 | 1003 | 1011 | 1011 | 918 | 946 | 972 | 743 | 841 | 870 |
| SONAR | 1012 | 1012 | 1012 | 1004 | 1012 | 1012 | 927 | 960 | 979 | 759 | 887 | 911 |
| Random | 1012 | 1011.67 | 1012 | 1003.67 | 1010.67 | 1011.33 | 884.33 | 926.67 | 948 | 691.33 | 770.67 | 809 |

Table 27: Statistics of example retrieval with different sentence embedding methods for $k$-shot settings ($k \in \{1, 5, 10\}$) on FLORES 200. For each setup, a value below 1012 suggests that some translations are either empty or written in an incorrect target language. The incorrect target language issue occurs more frequently than empty translations, except with `LLaMA 2 7B`, where a notable number of empty translations appear in the 1-shot setting for Swahili and Wolof. For random selection, we report the average across three runs.