Exploring In-context Example Generation for Machine Translation

Dohyun Lee[1] **Seungil Chad Lee**[1] **Chanwoo Yang**[2] **Yujin Baek**[1] **Jaegul Choo**[1]
[1]KAIST AI, [2]Jeonbuk National University,
{aiclaudev, seungil.lee, yujinbaek, jchoo}@kaist.ac.kr

Abstract

Large language models (LLMs) have demonstrated strong performance across various tasks, leveraging their exceptional in-context learning ability with only a few examples. Accordingly, the selection of optimal in-context examples has been actively studied in the field of machine translation. However, these studies presuppose the presence of a demonstration pool with human-annotated pairs, making them less applicable to low-resource languages where such an assumption is challenging to meet. To overcome this limitation, this paper explores the research direction of in-context example generation for machine translation. Specifically, we propose Demonstration Augmentation for Translation (DAT), a simple yet effective approach that generates example pairs without relying on any external resources. This method builds upon two prior criteria, *relevance* and *diversity*, which have been highlighted in previous work as key factors for in-context example selection. Through experiments and analysis on low-resource languages where human-annotated pairs are scarce, we show that DAT achieves superior translation quality compared to the baselines. Furthermore, we investigate the potential of progressively accumulating generated pairs during test time to build and reuse a demonstration pool. Our implementation is publicly available at https://github.com/aiclaudev/DAT.

1 Introduction

The recent emergence of large language models (LLMs) (Touvron et al., 2023a,b; OpenAI, 2023) and in-context learning (ICL) (Brown et al., 2020) has shifted the traditional paradigm of building task-specific models trained on large amounts of human-annotated data, which is costly to collect. The strength of ICL lies in its versatility, where they achieve outstanding performance across various tasks with just a few task-specific demonstrations (Yao et al., 2022; Wei et al., 2023). This reduces the training cost on large datasets and allows rapid adaptation to new domains or problems without extensive fine-tuning. Solving various tasks with a single LLM provides immense value to users seeking assistance in diverse contexts.

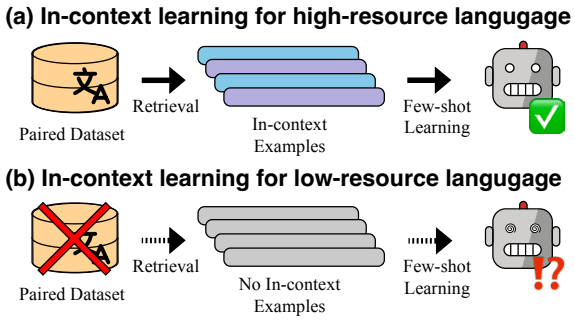


Figure 1: **Motivation.** (a) Previous works on LLM-based in-context learning for translation have primarily focused on selecting in-context examples from paired dataset in high-resource languages. (b) However, in the absence of a paired dataset for low-resource languages, how can in-context learning be applied?

LLMs have also played an increasingly prominent role in the field of machine translation (MT) due to their exceptional linguistic and reasoning capabilities (Moslem et al., 2022, 2023; Vilar et al., 2023; Jiao et al., 2023; Koneru et al., 2024; Xu et al., 2024). Notably, ICL has demonstrated high multilingual translation quality using only a few source-target pairs, driving advancements in retrieving optimal examples. Agrawal et al. (2023) proposed R-BM25, which initially selects the top bm25 candidates and reranks them using an n-gram recall strategy. Kumar et al. (2023) introduced a neural network trained to select pairs based on the multiple features, including semantic similarity and sentence length. These works have demonstrated outstanding performance in high-resource languages, retrieving pairs based on the scores between the given user query and source sentences in the demonstration pool, as shown in Figure 1 (a).

Findings of the Association for Computational Linguistics: ACL 2025, pages 26554–26568

However, these works on in-context example selection for MT may face challenges in low-resource languages. This stems from the fact that previous approaches rely on a critical assumption—namely, the availability of a large pool of human-annotated pairs—which may not hold for low-resource languages. For low-resource languages, obtaining a paired corpus for use as demonstrations is challenging due to the limited availability of public datasets and human annotators. This, in turn, poses a barrier that prevents low-resource languages from fully benefiting from the use of in-context examples, as illustrated in Figure 1 (b). Recently, El Mekki and Abdul-Mageed (2025) investigated leveraging LLMs to generate synthetic parallel data as a way to address this obstacle. However, this approach requires access to the vocabularies of both the source and target languages, as well as unlabeled sentences in the target language.

In this paper, we explore a research direction that aims to enable in-context learning for MT without the use of any external resources, instead drawing solely on the capabilities of the LLM itself. In pursuit of this goal, we introduce a simple yet effective method, Demonstration Augmentation for Translation (DAT), which utilizes the generative and linguistic capabilities of LLMs. This approach builds upon the intuitive prior criteria of *relevance* and *diversity*, which are inspired by previous works analyzing desirable in-context examples for MT (Cheng et al., 2022; Sia and Duh, 2023; Bouthors et al., 2024). To ensure these two criteria, we also utilize maximal marginal relevance (Carbonell and Goldstein, 1998).

Our experiment focuses on translating from English into low-resource languages—specifically Nepali, Khmer, Pashto, Zulu, and Swahili—for which the lack of extensive annotated datasets presents a realistic constraint. The results demonstrate the practicability of our easily applicable method in generating pairs that serve as in-context examples, providing valuable clues for user query translation. One more noteworthy point is that we observe a counterintuitive case where utilizing high-quality fixed pairs results in a severe performance degradation compared to the zero-shot approach. We investigate this phenomenon with a focus on the relevance between the source side of the pairs and the user queries. Lastly, we explore an extended method that incrementally accumulates the generated pairs and repurposes them through retrieval method such as R-BM25.

In summary, our contributions are as follows:

- To the best of our knowledge, this is the first work to explore in-context example generation specialized for MT without relying on any external resources, such as vocabularies or monolingual corpora.
- Experimentes show that DAT boosts the translation quality compared to other baselines, demonstrating its practicality for low-resource languages with scarce human-labeled pairs.
- Additional experiments demonstrate that high-quality fixed pairs in low-resource languages can act as noise and highlight DAT's potential for demonstration pool construction.

2 Related Work

2.1 In-context Learning

In-context learning (ICL) paradigm, originally proposed by Brown et al. (2020), enables LLMs (Touvron et al., 2023a,b; OpenAI, 2023; Dubey et al., 2024) to learn new tasks without any parameter updates by providing task-relevant input-output exemplars known as demonstrations (Liu et al., 2022). This paradigm facilitates incorporating human knowledge through task-specific examples into LLMs. It is often more effective than fine-tuning, allowing models to adapt to new cases with reduced data requirements (Mosbach et al., 2023). Previous works have introduced various strategies for constructing ICL prompts, highlighting that adjusting how demonstrations are composed can lead to more efficient solutions across various tasks (Zhao et al., 2021; Rubin et al., 2022; Hao et al., 2022; Cheng et al., 2023). Moreover, recent studies analyzing the factors influencing ICL performance have further supported its effectiveness (Min et al., 2022; Shin et al., 2022; Chan et al., 2022; Liu et al., 2022). Leveraging ICL, the ability of a single LLM to solve diverse tasks offers significant value in real-world applications.

2.2 Machine Translation using LLMs

Neural machine translation (NMT) models (NLLB Team et al., 2022) are trained with large amounts of high-quality parallel data, which is resource-intensive and costly. Consequently, numerous studies have been conducted on leveraging LLMs for machine translation, motivated by the data efficiency benefits offered by ICL. Extensive research has shown that leveraging zero-shot and few-shot

learning techniques achieves translation abilities that match or exceed the performance of traditional NMT models with just a minimal number of demonstrations (Lin et al., 2022; Chowdhery et al., 2023; Vilar et al., 2023; Zhang et al., 2023; Raunak et al., 2023a; Jiao et al., 2023). Recognizing the importance of effective demonstrations, further studies have focused on optimizing in-context example selection to improve the translation performance of LLMs through ICL, resulting in notable advancements in selection techniques (Agrawal et al., 2023; Kumar et al., 2023; Ji et al., 2024; Zebaze et al., 2025). However, these works rely on a large pool of human-annotated demonstrations, which can be impractical in real-world scenarios, especially for translation involving low-resource languages. To address this limitation, recently, El Mekki and Abdul-Mageed (2025) investigated leveraging LLMs to build a synthetic demonstration pool, but the approach requires access to lexical resources for both source and target languages, along with unlabeled data in the target language. In our method, we leverage only the linguistic capabilities of LLMs to generate in-context examples that enhance machine translation performance, marking the first attempt in the field.

2.3 Demonstration Augmentation

Demonstrations play a crucial role in ICL, as they significantly impact model performance by providing task-relevant examples that aid in solving new cases (Zhang et al., 2022; Lu et al., 2022; Liu et al., 2022; Bouthors et al., 2024). Research has moved beyond selecting high-quality examples, with growing interest in methods allowing LLMs to generate informative demonstrations autonomously (Kim et al., 2022; Lyu et al., 2023; Chen et al., 2023; Su et al., 2024). Li et al. (2024) confirmed that LLMs can achieve performance comparable to human-curated demonstrations by employing a self-reflective prompting strategy, illustrating that models can independently create examples that inform decision-making without the need for external, human-generated input. Our work builds on these advancements by exploring how leveraging the reasoning abilities of LLMs to generate source sentences and their translations produced by LLMs enhances translation performance. Especially, rather than using general approaches, we focus heavily on designing a more specific strategy for MT, eliminating reliance on human intervention or external data.

3 Method

Overview. We aims to generate pairs for in-context learning in the absence of human-annotated pairs. At test time, when a user provides a query q, LLM generates source-target pairs tailored to q and uses them as in-context examples. The overall flow of this method is shown in Figure 2.

3.1 Source-side of Pair Generation

To create source sentences $X = \{x_1, x_2, \cdots, x_m\}$ that provide valuable cues for translating q, we draw on previous research that explores the optimal in-context examples for MT (Cheng et al., 2022; Sia and Duh, 2023). These studies generally propose the following two priors:

- ***Relevance*** refers to the similarity between q and $x_i \in X$, which can include metrics such as n-gram overlap, edit distance, embedding similarity, and bm25 score.
- ***Diversity*** means the distinction between the retrieved examples, based on the intuition that if they are too similar to each other, they will provide redundant clues when translating q.

Generating X can be easily achieved by zero-shot prompting the LLM (source generator) with q and an instruction that incorporates the two priors. The simple instruction is illustrated in Figure 2 and the detailed prompt used for generating source sentences is presented in Figure 5.

3.2 Filtering using MMR

To ensure that the generated sentences satisfy the two prior conditions, we apply filtering and use only k examples. *Relevance* can be considered by a recall-based n-gram score R_n between q and x_i:

$$\mathrm{R}_n(q, x_i) = \frac{\sum_{\mathrm{ng} \in f_n(q) \cap f_n(x_i)} Count(\cdot)}{\sum_{\mathrm{ng} \in f_n(q)} Count(\cdot)} \quad (1)$$

$$\alpha(q, x_i) = \frac{1}{4} \sum_{n=1}^{4} \mathrm{R}_n(q, x_i), \quad (2)$$

where ng means n-gram and $f_n(\cdot)$ refers to the functions that convert a sentence into n-grams. Moreover, to promote diversity, we select examples using following equation inspired by Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) and Cheng et al. (2022).

$$\underset{x_i \in X \setminus X^*}{\operatorname{argmax}} [\alpha(q, x_i) - \frac{\lambda}{|X^*|} \sum_{x_j \in X^*} \alpha(x_j, x_i)], \quad (3)$$

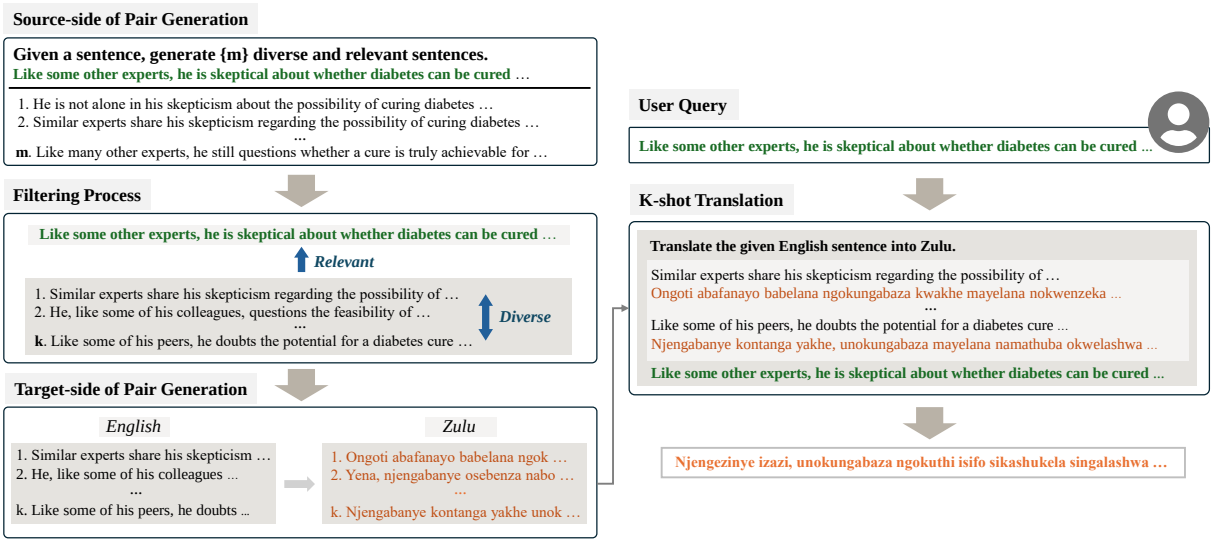


Figure 2: **An overview of our proposed method**. (1) Upon receiving a translation request for the user's query, the LLM generates m source-side sentences that satisfy both relevance and diversity constraints. (2) A relevant sentence that minimizes redundancy with previously chosen source sentences is iteratively selected and appended to the candidate pool. This process is iterated k times. (3) The LLM then translates each selected sentence, forming source-target pairs. (4) The final translation is produced through a few-shot learning framework, utilizing the generated pairs as in-context exemplars. A detailed explanation of our method is provided in Section 3.

where X^* is a set of already selected sentences and λ is a hyperparameter. MMR filtering process is detailed in Algorithm 1.

Algorithm 1 Filtering using MMR

Input: q, $X=\{x_i\}_{i=1}^{m}$, k ($<m$), λ
Output: Selected Sources $X^*=\{x_i^*\}_{i=1}^{k}$

procedure FILTERING(q, X, k, λ)
  $X^* \leftarrow \emptyset$
  while $|X^*| < k$ **do**
    for $x \in X \setminus X^*$ **do**
      $Relevance \leftarrow \alpha(q, x)$
      $Diversity \leftarrow \frac{-1}{|X^*|} \sum_{x_j \in X^*} \alpha(x_j, x)$
    end for
    $x^* \leftarrow \underset{x \in X \setminus X^*}{\operatorname{argmax}} (Relevance + \lambda Diversity)$
    $X^* \leftarrow X^* \cup \{x^*\}$
  end while
  return X^*
end procedure

3.3 Target-side of Pair Generation

After filtering, we need to generate translations for each $x^* \in X^*$. We have two options for translating the k source sentences: either using the LLM with zero-shot prompting or relying on the NMT model. LLM has acquired general knowledge across various domains through training on a vast pretraining dataset (Koneru et al., 2024) and excels at preserving semantic information (Hendy et al., 2023). For simplicity, we utilize LLM as our target generator.

3.4 Query Translation

By generating automatically without relying on any human-curated data, we now obtain k source-target pairs: $D^* = \{(x_i^*, \text{LLM}(x_i^*)\}_{i=1}^{k}$. Since these are tailored to the user query, they provide sufficient clues when translating the query. The query translator, an LLM, utilizes these demonstrations to perform in-context learning.

$$\hat{y} = \text{LLM}(I, D^*, q), \tag{4}$$

where $\hat{y}$ is translated from q and I refers to the instruction (e.g., `Translate a given <source language> sentence to <target language> sentence`). Figure 6 shows the detailed prompt.

4 Experimental Setup

4.1 Datasets and Languages

We benchmark the Flores dataset (Goyal et al., 2022), focusing on English and five low-resource languages: Nepali, Khmer, Pashto, Zulu, and Swahili. To evaluate performance, we conduct experiments on the devtest split, assessing the effectiveness of our approach in these language settings.

4.2 Evaluation Metrics

We utilize COMET[1] (Rei et al., 2022a), one of the most commonly used evaluation metrics. We also leverage reference-free COMET[2] (Rei et al., 2022b) to evaluate the quality of the in-context example in Table 2. For a more rigorous evaluation, we use afriCOMET[3] and reference-free afriCOMET[4] (Wang et al., 2024) for Zulu and Swahili, as afriCOMET is specialized for African languages. These metrics are designed to predict human judgments of translation quality.

4.3 Prompting Setup

Our proposed method is performed solely through zero-shot prompting, without relying on any pairs or examples. In our experiments, the number of in-context examples in few-shot prompting is fixed at 4. A source generator in our method initially generates 10 (m) sentences based on the user query with zero-shot prompting. Then filtering process remains 4 (k) sentences while considering relevance and diversity. The prompt template used in the whole experiments is provided in Figure 5 and 6.

5 Results and Analyses

5.1 Results on Low Resource Languages

Experimental Configuration Table 1 reports the COMET scores for translations from English into five low-resource languages. In real-world scenarios, these languages typically lack human-curated parallel corpora, which limits the feasibility of approaches such as in-context example selection and few-shot learning explored in previous work. As a workaround, one approach involves manually annotating a limited set of translation pairs and integrating them as fixed references during translation, which can serve as anchors and potentially enhance translation quality. This approach can also be applied to DAT, where these pairs are used during the target sentence generation process rather than when translating the test data.

No Fixed Pairs Setting Since this setting does not rely on human-annotated pairs, we investigate whether in-context examples, generated purely from the LLM's intrinsic capabilities, can serve as a catalyst for boosting translation quality beyond a zero-shot baseline. Our findings indicate that DAT has improved translation quality in most low-resource languages compared to zero-shot translation. Notably, for the Nepali, Llama-3.1-8B and 70B achieved performance gains of 2.8 and 1.3 points, respectively. This result shows that LLMs, relying solely on their inherent abilities without external information, enhance translation quality by using self-generated in-context examples. Consequently, it underscores their potential for generating low-resource language pairs.

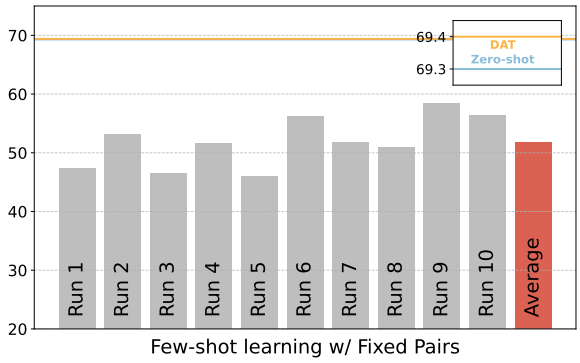


Figure 3: This figure illustrates English-to-Khmer translation results using fixed, high-quality human-aligned pairs. Each run is tested on 100 unique samples, with distinct fixed pair sets used across runs.

Fixed Pairs Setting A fixed pair set, while meticulously curated by humans to ensure high quality, may not always align with user queries requiring translation. Nevertheless, it can still facilitate performance improvement by offering linguistic cues related to language-specific grammar, syntactic structures, and other idiosyncratic features. This is exemplified by Llama-3.1-8B, which achieved a 3.8-point higher COMET score in few-shot method for the Nepali compared to the zero-shot method. Furthermore, DAT, which integrates a fixed human-curated data into the target sentence translation process, exhibits superior translation adequacy compared to the few-shot method across the majority of languages. This suggests that DAT enhances translation performance by dynamically generating semantically and syntactically aligned sentences while leveraging human-validated data to build high-fidelity translation pairs, resulting in greater adequacy and fluency.

The Backfire of Fixed Human Pair In this experiment, we observed a counterintuitive result: when translating from English to Khmer using Llama-3.1-70B, the few-shot approach led to a sig-

[1] Unbabel/wmt22-COMET-da
[2] Unbabel/wmt22-cometkiwi-da
[3] masakhane/africomet-stl
[4] masakhane/africomet-qe-stl

Fixed Pairs	Model	Method	Nepali	Khmer	Pashto	Zulu	Swahili
✘	*Llama-3.1-8B*	Zero-shot	72.1	62.0	53.9	**23.3***	60.6
		DAT	**74.9***	**64.4***	**54.6**	22.3	**61.8***
	Llama-3.1-70B	Zero-shot	79.8	**72.7**	67.5	37.8	72.9
		DAT	**81.1***	72.4	**68.3***	**38.3**	**73.4***
✔	*Llama-3.1-8B*	Few-shot	75.9	65.0	**57.9**	**24.7***	61.3
		DAT	**76.4**	**66.0**	57.3	23.3	**62.3***
	Llama-3.1-70B	Few-shot	80.6	51.1	65.7	38.9	71.6
		DAT	**81.5***	**52.9***	**68.5***	**39.2**	**72.7***

Table 1: The experimental results present COMET scores for translating English into five low-resource languages. Performance that surpasses the compared method is **bolded** for clarity, and * indicates statistically significant improvement at p=0.05, using the `compare-mt` library (Neubig et al., 2019). The "Fixed Pair" column specifies whether a fixed set of human-annotated pairs is utilized during the translation process. For a more detailed explanation of this setting, please refer to the experimental configuration in Section 5.1.

Method	Nepali				Khmer				Swahili			
	Relev.↑	Uni.↓	Qual.↑	COMET↑	Relev.↑	Uni.↓	Qual.↑	COMET↑	Relev.↑	Uni.↓	Qual.↑	COMET↑
Llama-3.1-70B												
Retrieval (src)	7.5	5.3	80.7	80.5	7.5	5.3	65.1	61.4	7.5	5.3	66.5	72.2
Fixed set (pair)	3.9	**2.8**	**89.4**	80.6	3.9	**2.8**	**85.6**	51.1	3.9	**2.8**	**77.2**	71.6
DAT	**25.9**	24.1	82.5	**81.1**	**25.9**	24.1	65.3	**72.4**	**25.9**	24.1	68.9	**73.4**

Table 2: This experiment presents the results of translating English into other low resource languages. Relevance (Relev.) measures the average n-gram overlap score between the user's query and the source side of an in-context example, while Uniformity (Uni.) evaluates the same averaged score among the source sides of different in-context examples. Quality (Qual.) is measured using reference-free COMET to evaluate the quality of a single pair. As a final point, COMET represents the score achieved when translating the user query with the given pairs. The best score in each column is highlighted in **bold**.

Method	Off-target Rate	# of Output Tokens
Zero-shot	0.0	241.7
Fixed Pairs	0.0	439.3
DAT	0.0	231.8

Table 3: These results present the off-target rate, indicating whether the output was translated into the correct language when translating from English to Khmer, along with the number of tokens in the generated sentences. We used `Google Translate` to identify the language.

nificant drop in translation quality—specifically, a 21.6-point decrease in COMET score compared to the zero-shot baseline. To assess the robustness of this finding, we ran 10 experiments using various fixed example pairs (Figure 3) and consistently found that few-shot translation underperformed relative to the zero-shot setting. As shown in Table 3, our analysis indicates that while the translations remained in the correct target language, incorporating fixed examples often resulted in abnormally long outputs. In some cases, the model repeatedly generated the same strings across different data points. However, our proposed method avoids these issues, which in turn leads to better translation performance. In Table 1, DAT without fixed pairs outperforms the few-shot approach in translation quality for all languages except Zulu. This raises a research question about whether it is better to use high-quality but fixed pairs that lack relevance to user queries, or moderate-quality pairs that more closely align to them. A thorough analysis of the underlying reasons is left for future work.

5.2 Quality of In-context Example

Experimental Configuration Previous studies have argued that in-context examples that are both similar to the user query and diverse from one another improve translation performance. In Table 2, we explore how the relevance of in-context examples to the user query, the uniformity among them—negatively correlated with diversity—and the intrinsic quality of example pairs affect translation performance. To investigate this, we consider the following methods: Retrieval, Fixed set, and DAT. Retrieval selects source sentences from a

monolingual pool using R-BM25 scores (Agrawal et al., 2023) and constructs pairs via an LLM, leveraging the accessibility of monolingual data. Fixed set relies on a predefined set of human-curated pairs, discussed in Table 1, as in-context examples. Meanwhile, DAT, our proposed method, generates pairs without relying on human-curated data, enabling a more autonomous approach.

Source-side of Pair Relevance and Uniformity exhibit the highest values in DAT and the lowest in the Fixed set. This disparity arises from the fact that both DAT and Retrieval dynamically generate or procure source sentences based on the user query, whereas the Fixed set operates independently of such adaptation. Meanwhile, Relevance and Uniformity appear to be interdependent, with an increase in Relevance generally leading to an increase in Uniformity. Though finding an appropriate balance between the two is important, Relevance is generally considered the higher-priority measure, as in-context examples that are diverse but unrelated to the user query are unlikely to be beneficial for translation. This is also supported by the DAT with the highest Relevance achieving a higher COMET score than other baselines.

Source-Target Pair While Relevance and Uniformity—focusing only on the source side of in-context examples—are important metrics, the overall quality of the pair itself is also a crucial factor. We evaluate this quality using the reference-free COMET score, reported in the Quality column. Despite the undeniable superiority of human-curated pairs in the Fixed set in terms of quality, our experimental results demonstrate that they yield lower translation performance compared to other baselines. This phenomenon can be attributed to their insufficient Relevance, which hinders their direct contribution to user query translation. This finding suggests that, rather than employing high-quality pairs that fail to ensure relevance to the user query, a more effective approach would be to generate or retrieve sentences with guaranteed relevance and then artificially construct their corresponding target sentences for use as in-context examples.

5.3 Ablation Study on MMR Filtering

Our method applies a filtering process to select only k sentences with high relevance and diversity from the initially generated m sentences. Then, we perform translation on the remaining k sentences to generate synthetic pairs by utilizing the LLM.

Method	m	k	Khmer	Pashto	Swahili
No Filtering$_4$	4	4	63.8	54.2	61.9
No Filtering$_{10}$	10	10	63.4	53.6	61.7
DAT	10	4	**64.4**	**54.6**	**62.3**

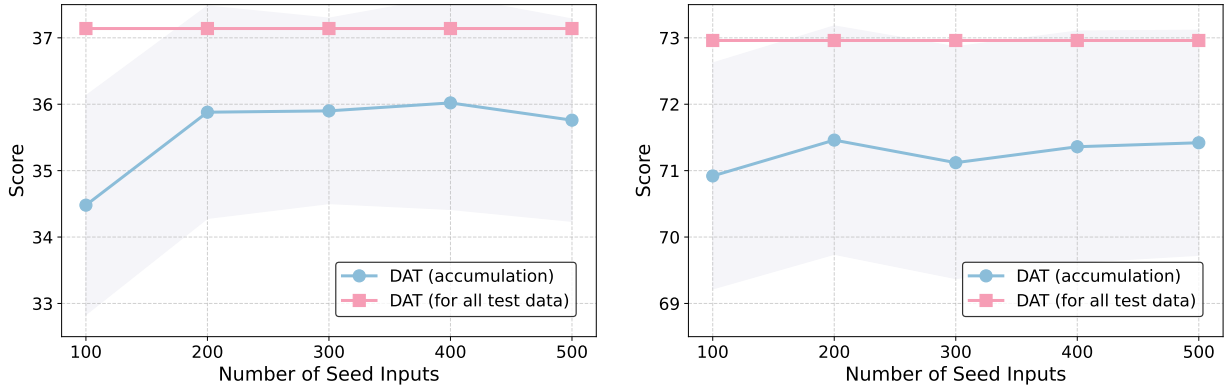
Table 4: This result illustrates the impact of m and k. m denotes the number of source sentences generated initially, whereas k signifies the final number of demonstrations after filtering process. This means that when m and k are the same, the filtering process is not applied. The best performances are in **bold**.

To verify the validity of process, we analyze the COMET scores when translating English into three low-resource languages in Table 4. It is noteworthy that No Filtering$_{10}$ showed inferior performance compared to No Filtering$_4$, despite using more in-context examples. This means that the additional 6 demonstrations acted as noise, degrading the translation quality. Furthermore, the results demonstrate that our approach—generating m source sentences with $m > k$ and then applying filtering—achieves superior performance compared to other methods. This validates the effectiveness of the filtering process and implies that using a small number of carefully selected demonstrations, which are filtered to better assist the user query, can lead to performance gains than utilizing 10 unfiltered demonstrations.

5.4 Accumulation Setting

Experimental Configuration In previous experiments, DAT demonstrated superior performance compared to baseline methods. However, generating in-context examples for every test input can be computationally expensive. To address this, we explore a more cost-efficient approach in this experiment. Specifically, we incrementally construct a demonstration pool by accumulating pairs generated for a subset of test inputs. Subsequently, we select in-context examples for translation based on their R-BM25 scores, thereby enhancing efficiency.

We partition the 1,012 test samples from the Flores dataset into two subsets of 500 and 512 instances. The first subset of 500 samples is designated as seed data, which we use to generate and accumulate pairs, thereby constructing the demonstration pool. Performance is then evaluated exclusively on the remaining 512 samples. Additionally, to investigate the impact of increasing seed data on performance, we progressively increase the number of seed data and assess the resulting improvements.



(a) From English into Zulu

(b) From English into Swahili

Figure 4: This experiment evaluates the accumulation setting when translating English into Zulu and Swahili using Llama-3.1-70B. The x-axis represents the number of test inputs used to construct the demonstration pool, while performance is assessed on a fixed set of 512 test samples, independent of the x value. To quantify variability, we further divide the test data into five distinct subsets and report the standard deviation.

Test Data	Reference	Method	In-context Example	Hypothesis
Prides are made up of one to three related adult males, along with as many as thirty females and cubs.	Makundi ya simba yanatengenezwa na waume kati ya mmoja na watatu wazima wanaohusiana, pamoja na wengi wa kike kama thelathini na mashibli.	Zero-shot		Vichwa vimeundwa na wanaume wazima 1 hadi 3 wenye uhusiano, pamoja na wanawake hadi 30 na watoto.
		DAT	A pride of lions can consist of up to thirty females and cubs, accompanied by one to three adult males. → Kundi la simba linaweza kuwa na hadi wanawake thelathini na watoto, wanaoambatana na wanaume wazima wawili hadi watatu.	Kundi la simba huundwa na wanaume wazima wawili hadi watatu wenye uhusiano, pamoja na wanawake hadi thelathini na watoto.

Table 5: This table shows a surface-level analysis when translating English to Swahili. Gray indicates cases where both Zero-shot and DAT produced the correct terms. Blue represents cases where only DAT generated the correct terms, demonstrating that DAT benefited from the self-generated in-context examples.

Result Figure 4 compares the COMET score of DAT, which performs few-shot learning by generating pairs for all test input, and DAT (Accumulation), which leverages a demonstration pool consisting of progressively accumulated pairs after a certain point. In both Zulu and Swahili, an increasing number of seed inputs exhibits a general trend of performance improvement. However, when employing a demonstration pool constructed from up to 500 data points, this approach does not fully reach the performance level of the method that generates pairs dynamically for each test instance. Nevertheless, we hypothesize that as more seed input is incorporated, a larger and more diverse demonstration pool can be constructed, ultimately enabling high-quality translations without the need for on-the-fly pair generation. Future research should explore optimal strategies for constructing and expanding such a pool, ensuring robust performance even in low-resource scenarios.

5.5 Surface-level Analysis

Table 5 illustrates the impact of in-context examples generated via DAT on translation outcomes, offering empirical evidence of their effectiveness. In this example, DAT successfully produces precise lexical choices that the Zero-shot approach fails to achieve. This is due to the alignment of generated in-context examples with the user query, which offers crucial contextual cues for precise translation.

6 Discussions

Hybrid Approach We employed an LLM to generate the target-side of in-context examples. While a model fine-tuned for a specific language pair—such as a traditional neural machine translation system—could produce higher-fidelity pairs, our focus is on improving translation quality purely through the inherent capabilities of the LLM. We therefore leave the exploration of such approaches as a promising direction for future work.

Reuse as a Training Dataset In Section 5.4, we explored a setting where the generated pairs are accumulated, allowing for retrieval and reuse. If the demonstration pool becomes large enough, it can serve as a training dataset for developing a translation-specialized model. Therefore, leveraging LLMs to generate translation pairs in low-resource scenarios is a crucial research direction, both for in-context learning and fine-tuning.

Post Editing Another line of research in translation using LLMs is post-editing, which focuses on refining the initial translation (Raunak et al., 2023b). Our research can be effectively combined with this method. In post-editing approaches, the quality of the initial translation is crucial. Applying DAT to generate the initial translation in scenarios where no in-context examples are available and then refining it presents a highly promising translation strategy.

7 Conclusions

In this paper, we explore an interesting research direction that leverages only LLMs to generate source-target pairs. Accordingly, we propose a simple yet effective method, DAT, which utilizes two priors and MMR to generate in-context examples. Experiments proved that DAT achieves superior translation quality for low-resource languages compared to baselines, without relying on any human-created resources. This highlights the potential of our method in scenarios where a demonstration pool is unavailable or fails to provide relevant translation examples. Furthermore, we investigated cases where fixed human pairs significantly underperform compared to zero-shot translation and explored the potential of an accumulation setting as a cost-efficient alternative.

8 Limitations

Translation from a Low-resource Language In this work, we focus on translating English into other low-resource languages. This is because most LLMs are primarily trained on English, allowing them to generate high-quality in-context examples on the source side. In contrast, translating in the opposite direction would heavily rely on the LLM's monolingual generation capabilities for low-resource languages. Therefore, this setting is not explored in this paper and remains an important challenge for future research.

Open-source LLMs Most open-source LLMs are primarily trained with a strong emphasis on English, limiting their effectiveness in handling a diverse set of languages. Given the scarcity of open-source models that offer robust multilingual support, we conduct our experiments using Llama-3.1, an open-source LLM designed to support a wide range of languages, including English.

Accumulation Setting In DAT with accumulation, the performance did not fully reach that of generating pairs for all test data. We aimed to create a larger demonstration pool to explore the point at which performance reaches that of generating pairs for all test data. However, this investigation could not be performed due to the high computational cost and is left for future work.

9 Ethical Considerations

We conducted our experiments using publicly available datasets and models, and the datasets do not involve any ethical concerns.

10 Acknowledgments

This work was supported by SAMSUNG Research, Samsung Electronics Co., Ltd., Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)), and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2025-00555621).

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.

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.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya K. Singh, Pierre H. Richemond, James L. McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Wei-Lin Chen, Cheng-Kuang Wu, Yun-Nung Chen, and Hsin-Hsi Chen. 2023. Self-icl: Zero-shot in-context learning with self-generated demonstrations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15651–15662. Association for Computational Linguistics.

Silei Cheng, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable. In *International Conference on Learning Representations (ICLR 23)*.

Xin Cheng, Shen Gao, Lemao Liu, Dongyan Zhao, and Rui Yan. 2022. Neural machine translation with contrastive translation memories. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3591–3601, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Sashank 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. 2023. Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1).

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 Ratanchandani, 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. *Preprint*, arXiv:2407.21783.

Abdellah El Mekki and Muhammad Abdul-Mageed. 2025. Effective self-mining of in-context examples for unsupervised machine translation with LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4229–4256, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. 2022. Structured prompting: Scaling in-context learning to 1, 000 examples. *CoRR*, abs/2212.06713.

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

Baijun Ji, Xiangyu Duan, Zhenyu Qiu, Tong Zhang, Junhui Li, Hao Yang, and Min Zhang. 2024. Submodular-based in-context example selection for llms-based machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 15398–15409. ELRA and ICCL.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt A good translator? A preliminary study. *CoRR*, abs/2301.08745.

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2022. Self-generated in-context learning: Leveraging autoregressive language models as a demonstration generator. *CoRR*, abs/2206.08082.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. Contextual refinement of translations: Large language models for sentence and document-level post-editing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 2711–2725. Association for Computational Linguistics.

Aswanth Kumar, Ratish Puduppully, Raj Dabre, and Anoop Kunchukuttan. 2023. CTQScorer: Combining multiple features for in-context example selection for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7736–7752, Singapore. Association for Computational Linguistics.

Rui Li, Guoyin Wang, and Jiwei Li. 2024. Are human-generated demonstrations necessary for in-context learning? The Twelfth International Conference on Learning Representations.

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.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8086–8098. Association for Computational Linguistics.

Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Z-ICL: zero-shot in-context learning with pseudo-demonstrations. 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 2304–2317. Association for Computational Linguistics.

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.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12284–12314. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2022. Domain-specific text generation for machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–30, Orlando, USA. Association for Machine Translation in the Americas.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 902–911. Association for Computational Linguistics.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, Xinyi Wang, and John Wieting. 2019. compare-mt: A tool for holistic comparison of language generation systems. *CoRR*, abs/1903.07926.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic 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.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023a. Do gpts produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1041–1050. Association for Computational Linguistics.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023b. Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. 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. 2022a. 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, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woo-Myoung Park, Jung-Woo Ha, and Nako Sung. 2022. On the effect of pretraining corpora on in-context learning by a large-scale language model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 5168–5186. Association for Computational Linguistics.

Suzanna Sia and Kevin Duh. 2023. In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 173–185, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Yi Su, Yunpeng Tai, Yixin Ji, Juntao Li, Bowen Yan, and Min Zhang. 2024. Demonstration augmentation for zero-shot in-context learning. *arXiv preprint arXiv:2406.01224*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George 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)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed, Ayinde Hassan, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-Azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Samuel Njoroge, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdullahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Raphael Ogbu, Sam Brian, Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoum Sari, Yao Lu, and Pontus Stenetorp. 2024. Afrimte and africomet: Enhancing comet to embrace under-resourced african languages. *Preprint*, arXiv:2311.09828.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. In-context example selection via similarity search improves low-resource machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*.

[System message]
You are a fluent assitant in {source_language}.
Given a sentence, generate {m} sentences based on the following criteria.

1. Relevance: Each sentence should be similar to the given sentence, but not identical to it.
2. Diversity: The generated sentences should be varied without duplicating each other.

The generated sentences must be separated from each other by \n and do not output any additional sentence such as explanation or reasoning in your response.

[User message]
{source sentence}

Figure 5: Prompt format used to generate source sentences with two criteria.

[System message]
You are a helpful assistant for translation.
Translate a given {source language} sentence to {target language} sentence.
Start the answer with [{target language}] and do not output any additional sentence such as explanation or reasoning in your response.

[User message]
[{source language}] {source sentence}

Figure 6: Prompt format used to translate a source sentence.

A Additional Information

To ensure the reproducibility of the experiments, we used a temperature of 0.1 during token decoding and the dataset statistics can be found in Table 6. Experiments is mainly conducted with RTX 3090 GPU. Zero-shot translation of the Flores devtest set using Llama-3.1-8B and 70B takes approximately 1 hour and 2 hours, respectively. We load Llama-3.1-8B and 70B from Hugging Face's Transformers (Wolf et al., 2020), with the LLaMA-3.1-70B model quantized to 4-bit. The fixed pair set used in the experiment consists of the 1st, 2nd, 3rd, and 4th data points from the dev set. We used ChatGPT to correct grammar.

train	dev	devtest
-	997	1012

Table 6: Data statistics of the Flores dataset (Goyal et al., 2022) used in all experiments.