SeEM: Selective Dictionary-based Generation

Hongyuan Lu[♠*], **Z.L.** [♠*], **Wai Lam**[♡]
[♠]FaceMind Research Asia
♡The Chinese University of Hong Kong
hongyuanlu@outlook.com

Abstract

While dictionary-based approaches have shown promise in applying Large Language Models (LLMs) to generation tasks such as Neural Machine Translation (NMT), existing studies overlook several key challenges. Most methods rely on full or rule-selected dictionaries for prompting, assume that all dictionary entries are effectively utilised by LLMs, and construct dictionaries from bilingual parallel corpora, each of which limits performance. We propose a new dictionary-based translation framework that addresses these issues by fine-tuning LLMs with LoRA to autonomously select salient dictionary entries, choosing translations with the highest dictionary utilisation, and introducing a Multilingual Lexicon Induced Dictionary that leverages multilingual parallel corpora to improve dictionary quality. Experiments on the FLORES benchmark across multiple languages demonstrate the superior performance of our framework.

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities across a wide range of applications. These capabilities include, but are not limited to chain-of-thought reasoning abilities (Wang et al., 2023; Wei et al., 2024), and multilingual neural machine translation (Zhu et al., 2024). More recently, dictionary-based methods effectively improve multilingual capabilities by incorporating multilingual word mappings (Lu et al., 2024) or bilingual word mappings (Ghazvininejad et al., 2023) into the prompt.

Although dictionary-based generative methods have shown promise, they still face at least three underexplored challenges. First, most methods use all matching dictionary without inspecting their importance, which causes potential redundant token usage. Second, LLMs may not follow the prompt instruction to fully utilize the provided dictionaries. Third, the dictionaries are usually automatically curated, and the construction process of dictionaries can be further improved. To address these issues, we propose a novel dictionary-based translation framework for LLMs called **Se**lective Dictionary-based Translation with **E**ndorsement Rate Vote and **M**ultilingual Lexicon Induced Dictionary (SeEM) to resolve the underexplored problems.

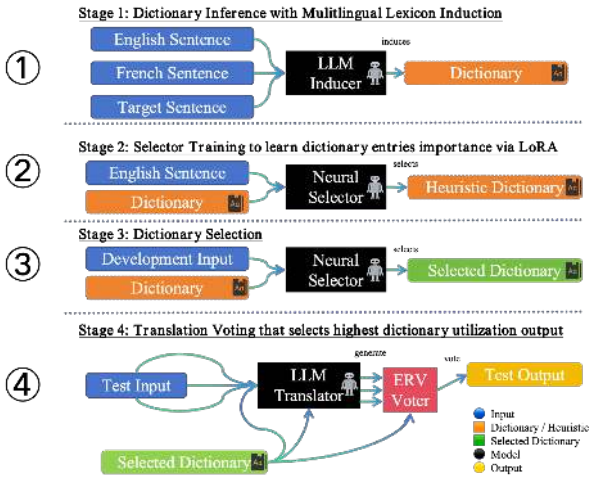


Figure 1: In the first stage of dictionary inference, we construct dictionaries using multilingual parallel sentences. These dictionaries are then used to train a neural dictionary selector, which identifies those yielding the best translation performance. The trained selector is subsequently employed to choose the optimal dictionaries. During the final translation, multiple different candidates are randomly generated with fixed hyperparameters, and the Endorsement Rate Vote (ERV) mechanism is applied to maximise dictionary utilisation among random translation generations. Note that the heuristic dictionary is constructed by searching the most useful dictionary entries. Such searches are done by adding and trying the dictionary and checking the quality with the reference. Here, the development set to select dictionaries and the test set are different to avoid test leakage.

Regarding the first challenge, Lu et al. (2024)

[*] Equal Contribution.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 4809–4830
November 6-10, 2026 ©2026 Association for Computational Linguistics

has shown that LLMs do not have to use all stop words for translation, and removing stop words can improve both token efficiency and the final performance on some languages. This is attributed to the fact that providing excessive content can degrade LLMs' performance (Shi et al., 2023). Given that the greedy string-match full dictionary has been dominantly used in this area, it can be linked with the supported capabilities of LLMs in recognising unknown and known knowledge to them (Amayuelas et al., 2024). To this end, SeEM proposes Selective Dictionary-based Translation, which fine-tunes LLMs to identify and select the most relevant dictionary entries. The training data is constructed by evaluating which dictionary subsets yield the greatest improvements in translation quality.

About the second challenge, we suspect that the dictionary usage is not guaranteed, because LLMs do not always follow human instructions given in the prompt (Lou et al., 2024). Our experiments further confirm this observation. To quantify dictionary usage, we introduce a novel metric, dictionary endorsement rate, which measures whether the generated output includes the desired dictionary entries. Then, to enhance dictionary usage, we propose Endorsement Rate Vote, a selection strategy that chooses the candidate translation with the highest dictionary endorsement rate among multiple generations.

The third challenge concerns the construction of dictionaries, which remains an underexplored and insufficiently evaluated process. One approach involves using LLMs for dictionary induction (Lu et al., 2024; Ghazvininejad et al., 2023); however, this method, commonly referred to as Bilingual Induction (Li et al., 2023) remains imperfect and leaves significant room for improvement. We propose a novel method called Multilingual Lexicon Induction to improve the accuracy in generating a dictionary by providing multilingual parallel sentences. This is motivated by the fact that multilingual data can enhance downstream bilingual tasks (Eisenschlos et al., 2019; Lai et al., 2022).

Overall, we propose SeEM, which is illustrated in Figure 1. SeEM clearly improves the dictionary-based translation method and is composed of the following three units:

- Selective Dictionary that fine-tunes LLMs to identify and select the most useful dictionary entries, outperforming existing selection baseline methods obviously.
- Endorsement Rate Vote that selects the generation with the highest dictionary endorsement rate among multiple candidates.
- Multilingual Lexicon Induction that leverages multilingual parallel corpora to build a dictionary with higher quality.

2 Prior Works

Neural Machine Translation on Large Language Models Prompting large language models (LLMs) for translation tasks is a popular research direction. Most existing methods focus on using LLMs for translation through simple instruction-based prompts such as 'Translate to {language_name}: text' (Brown et al., 2020; Lin et al., 2022; Le Scao et al., 2022; Zhang et al., 2022; Rezaeimanesh et al., 2025). In addition to those standard prompts, prior works have explored how the prompts can affect the translation quality(Reynolds and McDonell, 2021; Wang et al., 2023; Garcia and Firat, 2022). Furthermore, selecting samples similar to the source sentence for in-context learning is also useful for translation with LLMs (Agrawal et al., 2022; Vilar et al., 2022; Pei et al., 2025). LLMs have scaled their supported language for translation to at least hundreds of languages (Lu et al., 2024; Cui et al., 2025). However, their performance remains inferior to state-of-the-art NMT models on certain language pairs (NLLB-Team, 2022), and there is ongoing research on improving translation with LLMs.

Dictionary-based Translation on NMT Research on dictionary-based translation with LLMs (Ghazvininejad et al., 2023; Lu et al., 2024) is closely related to a prior research line using lexical constraints for the task of MT. The line can be categorised into either hard constraints (Hokamp and Liu, 2017; Post and Vilar, 2018) or soft constraints (Song et al., 2019; Dinu et al., 2019; Chen et al., 2021). The former are stricter than the latter, and methods are applied at different steps, such as decoding (Post and Vilar, 2018) and training (Dinu et al., 2019). Also, Zhang and Zong (2016) uses a bilingual dictionary to include rare or unseen words, which are lacking from the bilingual training data, to improve NMT models. Similarly, Arthur et al. (2016) improves the rare word translation using dictionaries along with attention vectors for estimating corresponding dictionary probabilities. Moreover, Hämäläinen and Alnajjar (2020)

also uses dictionaries to generate synthetic parallel training data to improve NMT models.

Dictionary-based Translation on Large Language Models The use of dictionaries in LLM-based translation is a relatively recent research direction. Ghazvininejad et al. (2023) uses bilingual dictionaries to improve translation on LLMs. Lu et al. (2024) further uses chained multilingual dictionaries to enhance machine translation with LLMs by leveraging intermediate auxiliary languages. Zhang et al. (2024) investigates the use of dictionaries even on unseen languages on LLMs. Nielsen et al. (2025) also investigates the use of string-match lexicons as dictionaries.

3 LLM Translation with SeEM

3.1 Translation with Large Language Models

We conduct experiments on the task of neural machine translation (NMT) using LLMs with access to external dictionaries. The objective is to translate source-language sentences into target-language equivalents while preserving their original meaning. The LLMs are optimised to maximise the following likelihood:

$$P\left(\hat{\boldsymbol{t}} \mid \boldsymbol{i}, \boldsymbol{s}, \boldsymbol{d}\right) = \prod_{j=1}^{\mathbb{T}} P\left(\hat{t}_j \mid \hat{t}_1, ..., \hat{t}_{j-1}, \boldsymbol{i}, \boldsymbol{s}, \boldsymbol{d}\right),$$

where $\mathbb{T}$ represents the length of the translation output and $\hat{t}_j$ represents the translation word at the position j that are to be generated. $\boldsymbol{s}$ represents the input sentence to be translated, $\boldsymbol{d}$ represents the dictionary entries that are provided to LLMs for reference. $\boldsymbol{i}$ represents the translation instruction. We use the following translation instruction:

> We have the following dictionaries: <dictionary>.
> Given the dictionaries, translate the following sentence from <source language> into <target language>: <source sentence>

3.2 Neural Dictionary Selection

However, using neural methods to select dictionary entries $\boldsymbol{d}$ remains an underexplored area. Most prior studies directly provide the entire string-matched dictionary to the prompt, regardless of its actual utility for translation. Yet, not all entries contribute equally to translation quality. To achieve a better trade-off between token efficiency and performance, it is desirable to measure which dictionaries are more important to the translators and translations, meaning which dictionaries can bring better translation performance. This further reduces the number of dictionary entries included in the prompt. We investigate neural dictionary selection methods to automatically identify dictionaries with higher token efficiency. We formally define this task using the following formula:

$$\hat{\mathrm{D}} = \mathrm{Model}(\mathrm{D}, \mathrm{Task}, \mathcal{V}),$$

where Model is a neural-based selection model, which is tuned to select a subset of dictionary entries $\hat{\mathrm{D}}$ from the full dictionary D for succeeding downstream Task. We define a dictionary size hyper-parameter $\mathcal{V}$ which is lower than the full dictionary size, and the selection model is optimised to seek a better subset of the dictionary which has a dictionary size of $\mathcal{V}$. The goal is that the dictionary should be selected to maximize the final downstream performance on the Task.

In order to construct the training data to obtain such a neural selection model, we need to know which dictionary entries are better and hence should be selected. We first construct dictionaries following prior work (Lu et al., 2024) using the following prompt with GPT-4o-mini:

> English:<en_sentence>
> <lang>:<lang_sentence>
> Please output the corresponding English-<lang> dictionary in each word. The format must be as follows: 'A' means 'B';

We feed each dictionary entry into translation:

> Translate the following sentence from lang to English.
> <origin_sentence>
> Use the provided dictionary to clarify or improve the translation of any misaligned words.
> - Here is a dictionary that you need to focus on: <dict>
> Note: only respond to me with the final English translation. Your output format is as follows:
> The refined translation is:

To construct the heuristic dictionary as in Figure 1 to train the neural selector, we try each dictionary

entry on translation on the dev set of FLORES, and we evaluate the combined improvements in BLEU and chrF++ scores, and use the dictionary entries that yield the highest gains as the training labels for training the neural dictionary selection model. Given an input sentence, the model predicts a subset of source-side dictionary entries, which are subsequently used for string matching to construct the final dictionary for translation.[1]

3.3 Endorsement Rate Vote

Since LLMs do not always follow instructions and do not always have to use the suggested dictionary entries, we propose a new metric called endorsement rate as follows:

$$E = \frac{|\hat{d}|}{|d|}, \tag{1}$$

where E represents the endorsement rate, $|\hat{d}|$ represents how many suggested target dictionary entries exist in the generated translation, and $|d|$ represents the total number of suggested dictionary entries. These metrics are calculated with the generated multiple translations as the inputs and are individually calculated, as in Step 4 in Figure 1.

Empirically, we observe that translation performance tends to improve as the endorsement rate increases, a trend that will be further demonstrated in the subsequent sections.

Therefore, SeEM proposes a novel method called Endorsement Rate Vote, which selects the one with the highest endorsement rate as the final translation out of multiple translations.

3.4 Multilingual Lexicon Induction

The dictionary construction stage is an interesting while understudied problem. Some prior works have formulated it as a bilingual lexicon induction problem, where bilingual parallel sentences are fed into LLMs to align words (Li et al., 2023).

However, Lu et al. (2024) demonstrated that leveraging high-resource languages as auxiliary information can enhance downstream task performance. Inspired by this finding, we adopt a similar approach for the task of bilingual lexicon induction and further introduce a new task: multilingual lexicon induction. In this task, large language models (LLMs) are provided with parallel sentences in multiple languages to facilitate more accurate word alignment. Specifically, we use the following prompt to construct the dictionary:

> English: <English_sensentence>;
> <lang>: <lang_sentence>;
> French: <French_sen>
> Please output the corresponding English-<lang>-French dictionary for each word. The format must be as follows: 'A' means 'B' means 'C';

Note that here, French is used as the auxiliary language. We only use the bilingual dictionary for the succeeding translation. We do not include the auxiliary language dictionary to make the comparison fair. As a result, any observed improvements can be attributed to enhanced word alignment.[2]

4 Experimental Setup

4.1 Dataset and Evaluation Metrics

We conduct our experiments on the task of machine translation using the FLORES benchmark (NLLB-Team, 2022). Specifically, we evaluate performance across 50 languages, each with approximately 1,000 test instances. For the metric, we report the chrF++ (Popović, 2015) and the BLEU (Papineni et al., 2002) provided by the sacreBLEU repository with default signature.[3] We use chrF++ signature of the parameters with nworde=6, ncorder=6, beta=2. We use BLEU signature of ngram=4, weights=(0.25, 0.25, 0.25, 0.25), smoothing=method1, smoothingfunction=SmoothingFunction().method1, tokenizer=nltkwordtokenize. We conduct evaluation with COMET scores using wmt22-comet-da[4] (Rei et al., 2020) throughout this paper.

[1] One finding in previous dictionary experiments is that the better the dictionary is, the better the translation will be (Lu et al., 2024). This paper assumes that the reverse side holds, and the better the translation is, the better the dictionary is. Since chrF++ (Popović, 2015) and the BLEU (Papineni et al., 2002) are typical evaluation metrics which have been proven to be useful to evaluate machine translation systems, we adopt them to evaluate and choose dictionaries. We use experiments to verify this assumption. To further verify that our method is better than simply choosing the words with the lowest frequency, we also compare with a relevant baseline (Lu et al., 2025)

[2] The orders are not always correct, but it is mostly correct (over 95%) and does not give a large impact to the results. The order problem was tested in Lu et al. (2024), and there are some matters, but they do not affect the performance much.

[3] https://github.com/mjpost/sacrebleu

[4] https://github.com/Unbabel/COMET

4.2 Models

LLMs For better reproducibility, our experiments are conducted on both closed-source (GPT-4o-mini) and open-source LLMs (LLaMA-3.1-8B (Grattafiori et al., 2024), DeepSeek-V3 (DeepSeek-AI et al., 2024)). DeepSeek-V3 is an Mixture-of-Expert LLM with 671B model parameters in total. We use it as a dictionary construction model. Both of these LLMs are widely known LLMs with promising multilingual capabilities. We also note that we use GPT-4o-mini and ChatGPT interchangeably throughout the paper, and we all refer to ChatGPT versioned GPT-4o-mini in our case.

Fine-tuning the Selector Model For the fine-tuning experiments aimed at selecting appropriate dictionary entries, we use Qwen2.5-7B-Instruct, an open-source large language model. We adopt LoRA fine-tuning for efficiency and parameter adaptation (Hu et al., 2022). The full list of fine-tuning hyperparameters is provided in the Appendix for reproducibility.

Multilingual Lexicon Induction To construct dictionaries, we utilise different models and compare their performance. Specifically, we evaluate both DeepSeek-V3 and GPT-4o-mini on this task and compare them against traditional bilingual lexicon induction methods for dictionary construction.

4.3 Baselines

We compare our proposed automatic selection method against strong baselines.

- **Vanilla Baseline** We directly ask LLMs to translate without any additional dictionary.
- **Random Baseline** We prompt LLMs to translate with randomly selected dictionary entries.
- **Noun Dictionary** Noun phrases may include named entities, which can be specialised terms that can be usually difficult to translate (Ugawa et al., 2018).
- **Adjective Dictionary** Adjectives are another type of word that could be important.
- **Noun and Adjective Dictionary** Combining both noun and adjective dictionaries for translation could be useful too.
- **Differ in Round-trip** Initially, we employ a baseline model without a dictionary to translate the source language into the target language and then back-translate it (Sennrich et al., 2016). The dictionary is derived from the differences between the round-trip translation and the original source sentences.
- **Differ in Translation** We start by using a baseline model without a dictionary to translate the source language into the target language. The differences between the translation and the original target sentences are then extracted to form the dictionary.
- **SLoW** This baseline selects low-frequency words for translation (Lu et al., 2025).
- **Selected Dictionary w/o Training** This baseline is a variant that uses a neural selector without any training.
- **Inverse Selected Dictionary w/o Training** This is a baseline with a neural-based selector directly used without training. We use those words not selected by the selector.
- **Selective Dictionary w/ Training** This is a component proposed in SeEM, with a neural model trained to select dictionaries.
- **Selective Dictionary w/f Training** This is a baseline that uses Selective Dictionary with filtered words that LLMs can translate in the development set. We use the LLMs to directly translate the input; the output is regarded as those words that the LLMs can translate.
- **Selective Dictionary w/f/b Training** This is a baseline that uses Selective Dictionary with filtered words that LLMs can translate in the development set. We also enrich this with back-translation in this baseline by back-translating the input and translating it into the output again to include more words.
- **SeEM w/o MLI** This is the SeEM framework we propose without Multilingual Lexicon Induction (MLI), but with the Selective Dictionary and Endorsement Rate Vote. Instead of MLI, we use Bilingual Lexicon Induction (BLI) for dictionary construction.

Note that we use GPT-4o-mini for dictionary construction by default, except in experiments specifically evaluating the effect of different models on dictionary construction.

5 Results

5.1 Main Results

Table 1 and Table 2 summarizes the results on SeEM compared the strong baseline SLoW on dif-

Direction	# improved > 5 points > 10 points > 15 points				# degraded > 5 points > 10 points		
En-X-LLaMA	48/50	47/48	33/48	15/48	2/50	1/2	1/2
En-X-ChatGPT	39/50	25/39	11/39	7/39	11/50	9/11	2/11
X-EN-LLaMA	50/50	49/50	37/50	19/50	0/50	0/0	0/0
X-EN-ChatGPT	48/50	34/48	19/48	10/48	2/50	1/2	1/2
X-X-LLaMA	45/50	39/45	30/45	14/45	5/50	3/5	2/5
X-X-ChatGPT	48/50	46/48	34/48	19/48	2/50	2/2	0/2

Table 1: Statistics of the changes in BLEU scores with SeEM compared to the strong baseline of SLOW on LLaMa and ChatGPT on the FLORES dataset. Most translation directions have been obviously improved. Note that we select the lowest scores here among SeEM variants for the comparison for each language pair.

Direction	# improved > 5 points > 10 points > 15 points				# degraded > 5 points > 10 points		
En-X-LLaMA	49/50	47/49	35/49	22/49	1/50	1/1	0/1
En-X-ChatGPT	36/50	23/36	15/36	7/36	14/50	11/14	3/14
X-EN-LLaMA	50/50	49/50	37/50	20/50	0/50	0/0	0/0
X-EN-ChatGPT	48/50	36/48	21/48	12/48	2/50	1/2	1/2
X-X-LLaMA	44/50	41/44	32/44	22/44	6/50	4/6	4/6
X-X-ChatGPT	48/50	46/48	34/48	22/48	2/50	2/2	0/2

Table 2: Statistics of the changes in BLEU scores with SeEM compared to the strong baseline of SLOW on LLaMa and ChatGPT on the FLORES dataset. Most translation directions have been obviously improved. Note that we select the lowest scores here among SeEM variants for the comparison for each language pair.

ferent language pairs on both LLaMA and ChatGPT, reported with both chrF and BLEU scores. Clearly, SeEM reports higher results on most results. For example, it reports 50/50 wins on X-EN-LLaMA. The worst results is on En-X-ChatGPT, with 36/50 wins. Still, this clearly shows SeEM's advantage as we are choosing the lowest score among SeEM variants for each language pair, which is a harder setting for SeEM.

For more baselines, Figure 8 in Appendix presents results on GPT-4o-mini, and Figure 9 shows results on LLaMA3. SeEM consistently outperforms all baselines—including its own ablated variants—across all evaluation settings. This demonstrates the effectiveness of SeEM. Among 50 languages, we have the following statistics. SeEM (with variants included) achieves 50/50 best results on both GPT and LLaMA across all translation directions. SeEM (with variants excluded, SeEM w/ GPT-MLI) yields 41/50 best scores on GPT and LLaMA on EX (English into other languages), 50/50 best scores on XE (other languages into English). It yields 50/50 best scores on GPT and 48/50 best scores on LLaMA when translating on XX (from other languages into other languages).

For the detailed number, we report the chrF++ scores with SeEM w/ GPT-MLI compared to the selected dictionary without training (we brief as 'baseline' in this paragraph). On the EX task, SeEM scores 73.27 vs. 55.04 on GPT and 63.13 vs. 38.51 on LLaMA. On XE, SeEM achieves 85.82 vs. 60.51 on GPT and 83.90 vs. 51.00 on LLaMA. On XX, SeEM scores 68.97 vs. 54.21 on GPT and 58.81 vs. 35.50 on LLaMA. In all cases, SeEM outperforms the baseline by a large margin.

We also report case studies in the Appendix along with their chrF++ scores on LLaMA. We observe that SeEM outperforms other baselines and variants clearly in terms of chrF++ scores.

5.2 Main Results on COMET Scores

SeEM gives the best results on all its variants and other baselines. This verifies the effectiveness of SeEM, compared to a competitive baseline using the selected dictionary without training. Among 50 languages, we have the following statistics. SeEM (with variants excluded, SeEM w/ GPT-MLI) yields 50/50 the best on GPT and 48/50 the best LLaMA on XE. It gives 38/50 on GPT and 44/50, the best on LLaMA on EX. It gives 35/50 on GPT and 49/50 the best on LLaMA on XX.

For the detailed number, we report here the COMET scores with SeEM w/ GPT-MLI compared to a selected dictionary without training (we brief

Direction	Model		Coverage		Avg. CHRF		Win Rate
	Translation	Dictionary	BLI	MLI	BLI	MLI	
EX	GPT-4o-mini	GPT	73.27%	**81.90%**	58.99	**73.27**	40/50
		DPK	56.36%	**75.29%**	63.04	**66.93**	49/50
	LLaMA	GPT	73.27%	**81.90%**	52.92	**63.13**	42/50
		DPK	56.36%	**75.29%**	56.20	**61.71**	49/50
XE	GPT-4o-mini	GPT	62.25%	**81.90%**	73.93	**85.82**	48/50
		DPK	52.04%	**79.35%**	70.13	**75.35**	50/50
	LLaMA	GPT	62.25%	**81.90%**	69.56	**83.90**	48/50
		DPK	52.04%	**79.35%**	66.50	**71.70**	50/50
XX	GPT-4o-mini	GPT	82.70%	**87.95%**	65.05	**68.97**	43/50
		DPK	80.29%	**85.48%**	62.06	**65.57**	40/50
	LLaMA	GPT	82.70%	**87.95%**	54.45	**58.81**	43/50
		DPK	80.29%	**85.48%**	50.77	**54.44**	42/50

Table 3: Ablation results on SeEM with MLI. We report both the dictionary coverage rate for the constructed dictionary against the ground-truth parallel sentences for dictionary construction, as well as the translation performance. This is calculated by the number of matching constructed dictionaries in the ground truth against the total number of words in the parallel sentences. The entries are treated as matching only if both the source word and target word are in the sentences. We experiment with the dictionary created with GPT and DeepSeek to be used for translation on GPT-4o-mini and LLaMA.

as 'baseline' in this paragraph). SeEM yields 72.57 versus baseline 67.34 on EX on GPT. SeEM yields 60.66 versus baseline 51.69 on EX on LLaMA. SeEM yields 81.70 versus baseline 74.69 on XE on GPT. SeEM yields 73.25 versus baseline 66.07 on XE on LLaMA. SeEM yields 67.96 versus baseline 62.45 on XX on GPT. SeEM yields 48.69 versus baseline 44.06 on XX on LLaMA. We also note that not all languages are supported by the COMET model we use. However, English are well-supported for all XE translation. Note that we release additional results on BLEU in the Appendix for space reasons.

5.3 Analysis on BLI versus MLI

Table 3 compares the performance of our proposed Multilingual Lexicon Induction (MLI) with the baseline Bilingual Lexicon Induction (BLI) in the task of dictionary construction for machine translation. Overall, MLI consistently outperforms BLI across all settings in terms of both translation quality and win rate.

For instance, on GPT-4o-mini translation, dictionaries constructed with MLI yield higher average chrF++ scores than those from BLI when using either GPT or DeepSeek (DPK) for dictionary creation. Specifically, MLI achieves 73.27 vs. 58.99 with GPT and 66.93 vs. 63.04 with DPK on the EX (English→X) task. Similar trends are observed on LLaMA, where MLI shows notable gains across directions compared to BLI.

Moreover, MLI delivers higher win rates in almost all scenarios, demonstrating its robustness and broader coverage. For example, on the XE (X→English) task, MLI achieves a perfect 50/50 win rate for both GPT-4o-mini and LLaMA when using DPK, outperforming BLI by a clear margin.

Due to space constraints, Table 3 presents only average scores and win rates. Complete results, including coverage rates and chrF++ scores for each language, are provided in Table 5, Table 6, and Table 7 in the Appendix.

The consistent improvements in both chrF++ scores and win rates across models, translation directions, and dictionary sources indicate that incorporating multilingual signals enables more accurate word alignment and better lexical coverage than traditional bilingual approaches. This consistently improves translation performance. This confirms that MLI is a superior strategy for enhancing dictionary-based translation in multilingual settings.

We also note that using GPT and DeepSeek for constructing dictionaries can give obviously different performance, where GPT seems to be better across GPT-4o-mini and LlaMA. This means that the dictionary construction model should be

Metric	Model	Method	EX	XE	XX
Util.	ChatGPT	**SeEM w/GPT-BLI w/o ERV**	0.7803	0.5572	0.7082
		SeEM w/GPT-BLI w/ ERV	**0.8469**	**0.6514**	**0.7576**
	LLaMA	**SeEM w/GPT-BLI w/o ERV**	0.6514	0.5376	0.6233
		SeEM w/GPT-BLI w/ ERV	**0.8118**	**0.6565**	**0.6490**
Util. Winning Rate	ChatGPT	**SeEM w/GPT-BLI w/o ERV**	4/50	0/50	5/50
		SeEM w/GPT-BLI w/ ERV	**46/50**	**50/50**	**45/50**
	LLaMA	**SeEM w/GPT-BLI w/o ERV**	1/50	0/50	9/50
		SeEM w/GPT-BLI w/ ERV	**49/50**	**50/50**	**41/50**
chrF++	ChatGPT	**SeEM w/GPT-BLI w/o ERV**	57.43	63.35	56.44
		SeEM w/GPT-BLI w/ ERV	**58.99**	**73.93**	**65.05**
	LLaMA	**SeEM w/GPT-BLI w/o ERV**	47.50	57.67	45.32
		SeEM w/GPT-BLI w/ ERV	**52.92**	**69.56**	**54.45**
chrF++ Winning Rate	ChatGPT	**SeEM w/GPT-BLI w/o ERV**	16/50	0/50	0/50
		SeEM w/GPT-BLI w/ ERV	**34/50**	**50/50**	**50/50**
	LLaMA	**SeEM w/GPT-BLI w/o ERV**	8/50	0/50	4/50
		SeEM w/GPT-BLI w/ ERV	**42/50**	**50/50**	**46/50**

Table 4: Ablation results on SeEM with ERV. Util. represents the dictionary coverage rate for the translation. ERV represents endorsement rate vote. This table represents the Util. and chrf statistics for SeEM with ERV and without ERV. The winning rate is reported for 50 languages from the FLORES dataset. Overall, we can see that the Util. are usually improved with ERV, and such improvements consistently exist with the improvements in chrF++. This aligns with our initial expectations to use ERV to increase Util. to improve the final translation performance.

selected carefully to maximise the performance.

For space reasons, Table 6 in Appendix presents the analysis on matched entries' frequency with BLI and MLI. The lower avearged score on frequency may suggests the better ability of MLI in matching rare words.

5.4 Analysis on ERV and latency

Table 4 reports the results of an ablation study evaluating the impact of the Endorsement Rate Vote (ERV) mechanism within the SeEM framework. Two key observations emerge from the comparison between SeEM w/ and w/o ERV: i) Improved Dictionary Utilization: Incorporating ERV consistently leads to higher dictionary coverage across all translation directions (EX, XE, XX) and both models (ChatGPT and LLaMA). For example, on GPT-based EX translation, dictionary utilization improves from 0.7803 to 0.8469 with ERV. ii) Enhanced Translation Quality: ERV also brings consistent gains in chrF++ scores. For instance, chrF++ increases from 57.43 to 58.99 on GPT EX, and from 47.50 to 52.92 on LLaMA EX.

For space reasons, we leave latency analysis in Appendix D, which shows acceptable latency (+10%) with SeEM on the best setting.

6 Conclusions

We propose a novel dictionary-based translation framework for LLMs called **Se**lective Dictionary-based Translation with **E**ndorsement Rate Vote and **M**ultilingual Lexicon Induced Dictionary (SeEM). SeEM is composed of three units. The first unit uses LoRA to train neural selectors to obtain the final translations. The second unit uses the Endorsement Rate Vote to enhance the instruction following capabilities of LLMs, which selects the translations that best use the dictionary entries. The third unit proposes Multilingual Lexicon Induction that improves the dictionary construction. Experiments on the FLORES dataset with EX, XE, and XX directions suggest the effectiveness of SeEM.

Limitations

Current methods relevant to model training are limited to LoRA experiments with small LLMs, which is due to resource limitations. Yet, such circumstances simulates a model usual use case, which reduces the user entrance barrier.

Ethical Statement

We honour and support the ACL ARR Code of Ethics. The datasets used in this work are well-known and widely used, and the dataset preprocessing does not make use of any external textual resource. In our view, there is no known ethical issue. End-to-end pre-trained LLMs are also used, which are subject to generating offensive content. But the above-mentioned issues are widely known to commonly exist for these models. Any content generated does not reflect the view of the authors.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context Examples Selection for Machine Translation. *arXiv e-prints*, arXiv:2212.02437.

Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Yang Wang. 2024. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6416–6432, Bangkok, Thailand. Association for Computational Linguistics.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. 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.

Guanhua Chen, Yun Chen, Yong Wang, and Victor O. K. Li. 2021. Lexical-constraint-aware neural machine translation via data augmentation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi

Gao, and Zizheng Pan. 2024. DeepSeek-V3 Technical Report. *arXiv e-prints*, arXiv:2412.19437.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Julian Eisenschlos, Sebastian Ruder, Piotr Czapla, Marcin Kadras, Sylvain Gugger, and Jeremy Howard. 2019. MultiFiT: Efficient multi-lingual language model fine-tuning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5702–5707, Hong Kong, China. Association for Computational Linguistics.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *arXiv e-prints*, arXiv:2202.11822.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based Phrase-level Prompting of Large Language Models for Machine Translation. *arXiv e-prints*, arXiv:2302.07856.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, 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, Danny Wyatt, 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, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, 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, Jack Zhang, 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, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, 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, Maria Tsimpoukelli, 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, Ning Zhang, 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, Rohan Maheswari, 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, and Tobias Speckbacher. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, arXiv:2407.21783.

Mika Hämäläinen and Khalid Alnajjar. 2020. A template based approach for training nmt for low-resource uralic languages - a pilot with finnish. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI '19, page 520–525, New York, NY, USA. Association for Computing Machinery.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2022. Multilingual pre-training with language and task adaptation for multilingual text style transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 262–271, Dublin, Ireland. Association for Computational Linguistics.

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, 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. 2022. BLOOM: A 176B-Parameter

Open-Access Multilingual Language Model. *arXiv e-prints*, arXiv:2211.05100.

Yaoyiran Li, Anna Korhonen, and Ivan Vulić. 2023. On bilingual lexicon induction with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9577–9599, Singapore. Association for Computational Linguistics.

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.

Renze Lou, Kai Zhang, and Wenpeng Yin. 2024. Large language model instruction following: A survey of progresses and challenges. *Computational Linguistics*, 50(3):1053–1095.

Hongyuan Lu, Zixuan Li, Zefan Zhang, and Wai Lam. 2025. SLoW: Select low-frequency words! automatic dictionary selection for translation on large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 898–913, Suzhou, China. Association for Computational Linguistics.

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

Elizabeth Nielsen, Isaac Rayburn Caswell, Jiaming Luo, and Colin Cherry. 2025. Alligators all around: Mitigating lexical confusion in low-resource machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 206–221, Albuquerque, New Mexico. Association for Computational Linguistics.

NLLB-Team. 2022. No language left behind: Scaling human-centered machine translation.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schütze. 2025. Understanding In-Context Machine Translation for Low-Resource Languages: A Case Study on Manchu. *arXiv e-prints*, arXiv:2502.11862.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. 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.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Sara Rezaeimanesh, Faezeh Hosseini, and Yadollah Yaghoobzadeh. 2025. Large language models for Persian-English idiom translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7974–7985, Albuquerque, New Mexico. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

Volume 1 (Long and Short Papers), pages 449–459, Minneapolis, Minnesota. ACL.

Arata Ugawa, Akihiro Tamura, Takashi Ninomiya, Hiroya Takamura, and Manabu Okumura. 2018. Neural machine translation incorporating named entity. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3240–3250, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting PaLM for Translation: Assessing Strategies and Performance. *arXiv e-prints*, arXiv:2211.09102.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. *arXiv e-prints*, arXiv:2303.04048.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024. Teaching large language models an unseen language on the fly. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8783–8800, Bangkok, Thailand. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Bridging Neural Machine Translation and Bilingual Dictionaries. *arXiv e-prints*, arXiv:1610.07272.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. *arXiv e-prints*, arXiv:2205.01068.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Appendix

A LoRA Finetuning Hyperparameters

Hyperparameter	Value
quantization_bit	4
stage	sft
do_train	true
finetuning_type	lora
lora_target	all
template	qwen
cutoff_len	1024
max_samples	3000
overwrite_cache	true
preprocessing_num_workers	16
logging_steps	10
save_steps	500
per_device_train_batch_size	1
gradient_accumulation_steps	8
learning_rate	1.0e-4
num_train_epochs	10.0
lr_scheduler_type	cosine
warmup_ratio	0.1
bf16	true

Table 5: A list of hyperparameters used in our finetuning experiments.

B Additional Results on BLEU Scores

We present the BLEU results on GPT-4o-mini in Figure 10 (Appendix), and on LLaMA in Figure 11. Overall, SeEM consistently achieves the best performance across all its variants and against all baselines, demonstrating its effectiveness compared to a strong baseline that uses a selected dictionary without training. Across 50 languages, the summary statistics are as follows: SeEM (including all variants) achieves 50/50 best results on both GPT and LLaMA for XE (other languages to English). For EX (English to other languages), it achieves 50/50 on GPT and 40/50 on LLaMA. For XX (non-English to non-English), it reaches 50/50 on GPT and 49/50 on LLaMA. SeEM (with variants excluded, SeEM without variants (i.e., SeEM w/ GPT-MLI) achieves: 40/50 best scores on GPT and 38/50 on LLaMA for EX; 49/50 on GPT and 50/50 on LLaMA for XE; 50/50 on GPT and 48/50 on LLaMA for XX.

For the detailed number, we report here the BLEU scores with SeEM w/ GPT-MLI compared to selected dictionary without training (we brief as 'baseline' in this paragraph). SeEM yields 43.52 versus baseline 26.04 on EX on GPT. SeEM yields 34.11 versus baseline 11.69 on EX on LLaMA. SeEM yields 55.69 versus baseline 31.53 on XE on GPT. SeEM yields 53.95 versus baseline 22.59 on XE on LLaMA. SeEM yields 39.72 versus baseline 25.55 on XX on GPT. SeEM yields 30.12 versus baseline 8.37 on XX on LLaMA.

C Frequency Analysis on MLI and BLI

Category	BLI	MLI
Very Rare (zipf<2)	2.7%	3.4%
Rare (zipf 2-4)	11.9%	12.3%
Common (zipf 4-6)	46.2%	46.5%
High-freq (zipf > 6)	39.3%	37.8%
Avg entries per sentence	22.27	21.7
Mean Zipf frequency	5.44	5.37

Table 6: The frequency analysis on MLI and BLI, constructed with 1,012 samples from FLORES dev set.

D Cost Analysis

Method	Latency (s)	Overhead
Vanilla LLM (0-shot)	2.06	—
Selector only (batchsize=1, single)	0.996	—
SeEM w/o ERV (selector+translation)	3.01 [0.996+2.014]	+46%
SeEM w/ ERV (n=5)(selector+translation)	3.03 [0.996+2.034]	+47%
Selector only (bs=8, batched)	0.181	—
Selector only (bs=16, batched)	0.118	—
Selector only (bs=24, batched)	0.073	—

Table 7: We presents the analysis here for the SeEM inference. When we do batching for ERV (n=5) vs w/o ERV, it adds +0.02s more time. If we do selector speedup via batching, it does 0.996s → 0.181s (82% reduction) on SeEM. So for the best setting, the addtional averaged latency is 0.201s with SeEM on our setting.

The translation model (GPT-4o-mini, etc.) is called via API. GPU: single NVIDIA 4090 Qwen2.5-7B-Instruct serves as the local Selector after fine-tuning. Analysis: The Selector is the primary source of additional overhead in SeEM, but this step can be batch-processed by increasing the batchsize. ERV operates in parallel, so the total latency is nearly equivalent to that of a single translation. The average per-sentence latency is shown in Table 7: (In single-sentence testing, the SeEM stage first runs the Selector and then the Translation; in actual processing, the Selector is batch-processed first, and then the dictionary entries are incorporated into the translation task.)

E Cases

Source Sentence (eng_Latn): USA Gymnastics supports an independent investigation that may shine light on how abuse of the proportion described so courageously by the survivors of Larry Nassar could have gone undetected for so long and embraces any necessary and appropriate changes.		
Target Sentence (bam_Latn): USA Jiminasitiki be dƐmƐdo ɲinnin kƐrƐnkƐrƐnnen dɔ ma min be na se ka tiɲƐ bange cogo kan min ka damatƐmƐwale lakalira mabalolenw fƐ ni timinandiya ye cogo min ni Larry Nassar sera ka tƐmƐ ka mƐƐ kasɔrɔ a ma ye ni ka se ka falenniw kƐ minw tun ka kan ka kƐ.		
Model Name	Chrf	Translation
Vanilla	20.53	USA Gymnastics bɛ bɔ kɛ fɛ ka fɔlɔ kɛla kan bɛɛ kɛ bɛɛya la, ni ye faamuya jɔya ka taa ka ladilikan kɛ fɔlɔ, o ye Larry Nassar ye sigilanw ye ka kɔrɔfɛɛ min y'a fɔ ka yɛrɛyɛrɛ kɛ, a ye sigira kelen don o sisan kɔ.
Random	33.52	USA Jiminasitiki bɛ bɔ dɔ ɲinnin kƐrƐnkƐrƐnnen kɛ bɛ bange cogo kan, i fƐrƐ ye mɛn fɔ dɔ min kelen ye fƐ ko survivors fƐ Larry Nassar kɛ la ɲa bɔ ko kɛ kɔrɔ kɔnɔ ye fɔ, a bɛ bɔ minw fɛnw kɛ a bɔ yera fƐrƐfɛrƐ fƐ ye.
Noun	33.02	USA Jiminasitiki bɛ bɛɛ ka dɔ kɛ kɔfɛ, k'a bange don kɔrɔya, a la timinandiya ye, o ye Larry Nassar ye kasɔrɔ a ma bɛɛ ka fɔ i ka dɔ bɔ, ni o bɛ ka taa fitila bɛɛ, ni USA Jiminasitiki bɛ ka se ka walima ni tile dɛmɛ kɛ kɔfɛ ye.
Adjective	33.21	USA Gymnastics bɛ kɛrɛnkɛrɛnnen bɛɛ ka fɛ ka taa, kɔrɔ lakalira mabalolenw kɛra, ni Larry Nassar sera ka fɔ a la, o bɛ kɛ kɛ minw tun ka kan ka kɛ i ka taa don, ni bɛ bɔ fo tun ka kan ka kɛ ni ka bɛɛ ka.
Noun and Adjective	37.33	Jiminasitiki Amerika be kƐrƐnkƐrƐnnen dɔ da yira bange la, ka fɔ ko Larry Nassar sera kasɔrɔ a ma ye tuguya fɛ kɔrɔbɔ ye bɔ ka taa, a ni ka tun ka kan ka kƐ ni ka fɔlɔ fɛ ka kɛ ani ka fɛn bɛɛ bɛɛ.
Differ in round-trip	24.95	USA Gymnastics be dɛmɛdo kɛrɛnkɛrɛnnen dɔ la, ni o bɛ tugu light kan ka fɔ ko kɔrɔbaya kɛ jan ni Larry Nassar danbiliw ye bɛɛ fɔ ka cɛya falenniw i ni tile dɔ, ani bɛ dɛmɛ minw ka ni ka gɛlɛya ladon bɛɛ.
Differ in translation	33.11	USA Gymnastics bɛ dƐmƐdo kƐrƐnkƐrƐnnen ɲinnin, bɛ tiɲƐ ka fɛn ye, ka tƐmƐ Larry Nassar fanga bɔ yɔrɔ min bɛɛ ka fɔ ko yɛrɛ kɛlɛya la, a bɛ ta kɔrɔ kɔrɔ, ani bɛ dɛmɛ minw tun ka kan ka kƐ ani min bɛ tilen ka bɔnbɔ.
SLoW	40.44	USA Jiminasitiki bɛ ka kƐrƐnkƐrƐnnen dɔ dɛmɛ, k'a tiɲƐ bange ka kɛ kƐ minw tun ka kan ka kƐ Larry Nassar sera ye fili-fili ye, ani bɛ ka se ka minw fɛnw b'a la ani fɛ ka fɛ.
Selected dictionary w/o training	28.13	USA Gymnastics bɛ dɛmɛdo kɛrɛnkɛrɛnnen dɔ la, min bɛ siran ka taa kɔnɔya la lakalira mabalolenw bɛɛ, o bɛɛ ye kasɔrɔ a ma ye kɔrɔbɔya fɛ, ani bɛ bɛ dɛmɛ falenniw bɛɛ min bɛ bɛ se ka bonya ani ka taa kɔnɔya la.
Inverse selected dictionary w/o training	32.66	USA Jiminasitiki bɛ bɔ an ɲinnin min bɛ fisa don kɔfɛ la, k'a fɔ ko abusɛ fɛ bɛ kɛra, ni ko Larry Nassar la ka taaɲɛw ye ka fɔ ko, i b'a don waati juguman kelen tuma la, ni bɛ ka se ka yera muso tuma la, ni USA Jiminasitiki bɛ ka se ka fili kelen ni min bɛ nyɛ ka bonya kɛ.
Selective dictionary w/ training	53.44	USA Jiminasitiki be dƐmƐdo dɔ kƐrƐnkƐrƐnnen min bange ka fɛ min ka damatƐmƐwale lakalira mabalolenw fƐ ka mɛn bɔ min ka kɛnɛya ka fɔ Larry Nassar bɛɛ kɛ minw tun ka kan ka kƐ kɛnɛya ka fɔ bɔ fɛ ka taa la, o be dɔɔni ka fɛrɛ kɛ mɔgɔkunw ye ka fɔ ka mɛn yera ni ka fɛrɛ kɛ sisan kɛ.
Selective dictionary w/f training	58.02	USA Jiminasitiki be dƐmƐdo kƐrƐnkƐrƐnnen dɔ fƐ ka bange ka min ka damatƐmƐwale lakalira mabalolenw fƐ ka mƐƐ kan ka kƐ kƐ minw tun ka kan ka kƐ ka taa fɔ Larry Nassar fili kƐlenw ye, ka mɛn bɛɛ ta fɛn ni falenniw fɛrɛfɛrɛ ni sisanw bɛɛ ye.
Selective dictionary w/f/b training	57.61	USA Jiminasitiki be dƐmƐdo kƐrƐnkƐrƐnnen dɔ fƐ bange bɛɛ min ka damatƐmƐwale lakalira mabalolenw fƐ kɛnɛya ka kasɔrɔ a ma Larry Nassar fila kƐ minw tun ka kan ka kƐ kɔrɔ la, ni bɛ fɔ falenniw ni tɛmɛ ni sɛbɛn fƐ.
SeEM w/DSK-BLI	50.01	USA Jiminasitiki be dƐmƐdo kƐrƐnkƐrƐn dɔ kƐ bange cogo kan la lakalira mabalolenw fƐ fɛ ka fɔlɔ kasɔrɔ a ma Larry Nassar ye dɔ ko ni baro kɛnɛ tƐmƐ ka tƐmƐ ka kan ka kɔrɔ, ni be dƐmƐdo falenniw kɛ ni fɛ ka fɔ.
SeEM w/GPT-BLI	57.68	USA Jiminasitiki be dƐmƐdo ɲinnin kƐrƐnkƐrƐnnen dɔ ma min be na se ka tiɲƐ bange cogo kan min ka damatƐmƐwale lakalira mabalolenw fƐ ni timinandiya ye cogo min ni Larry Nassar sera ka tƐmƐ ka mƐƐ kasɔrɔ a ma ye ni ka se ka falenniw kƐ minw tun ka kan ka kƐ.
SeEM w/DSK-MLI	58.69	USA Jiminasitiki be dƐmƐdo kƐrƐnkƐrƐnnen dɔ fɛ bange ka min ka damatƐmƐwale lakalira mabalolenw fɛ, kɛ kasɔrɔ a ma Larry Nassar kɛrɛ ye a ka fɔlɔ cogo dɔ ka kƐ minw tun ka kan ka kƐ, ani ka falenniw fɛlɛ ka bɔra.
SeEM w/GPT-MLI	65.49	USA Jiminasitiki be dƐmƐdo kƐrƐnkƐrƐnnen dɔ la, k'a bange cogo kan min ka damatƐmƐwale lakalira mabalolenw fƐ, ni kasɔrɔ a ma ye Larry Nassar la min ye a fɔ kƐlƐya la, ni kƐ minw tun ka kan ka kƐ i ni juman fƐ. A be na falenniw bɛɛ min bɛ se ni min bɛ to.

Figure 2: Case study on translating following our proposed framework on EX

Source Sentence (Modern Standard Arabic (Romanized)): Wa yushakkek al-doktor Ehud Ur, mithil ba3edh al-5ubara2 al-a5arin, fi imkaniyat ashifa2 min marad as-soukari, mushiran ila anna hadhihi annata2ij laysa laha silaton bil ashkhas al-musabin bi marad as-soukari min an-naw3i al-awwal.		
Target Sentence (eng_Latn): Like some other experts, he is skeptical about whether diabetes can be cured, noting that these findings have no relevance to people who already have Type 1 diabetes.		
Model Name	Chrf	Translation
Vanilla	42.14	Dr. Ehud Ur, like some other experts, doubts the possibility of a cure for diabetes, pointing out that these results are not related to individuals with type 1 diabetes.
Random	47.96	And Dr. Ehud Ur, like some other experts, doubts the possibility of cures for diabetes, indicating that these results have no connection to the people affected by type 1 diabetes.
Noun	62.08	And Dr. Ehud Ur, like some other experts, doubts the possibility of curing diabetes, noting that these findings have no relevance to people affected by type 1 diabetes.
Adjective	53.84	Dr. Ehud Ur, like some other experts, is skeptical about the possibility of a cure for diabetes, pointing out that these results have no connection to people with type 1 diabetes.'
Noun and Adjective	59.95	Dr. Ehud Ur, like some other experts, doubts the possibility of a cure for diabetes, noting that these findings have no relevance to patients with type 1 diabetes.
Differ in round-trip	54.23	And Dr. Ehud Ur, like some other experts, doubts the possibility of a cure for diabetes, pointing out that these results have no relevance to people with type 1 diabetes.
Differ in translation	57.64	Dr. Ehud Ur, like some other experts, is skeptical about the possibility of a cure for diabetes, noting that these results have no connection to people with type 1 diabetes.
SLoW	68.87	And Dr. Ehud Ur, like some other experts, is skeptical about the possibility of being cured from diabetes, noting that these findings have no relevance to people suffering from type 1 diabetes.
Selected dictionary w/o training	54.25	And Dr. Ehud Ur, like some other experts, doubts the possibility of a cure for diabetes, pointing out that these findings have no relation to people with type 1 diabetes.
Inverse selected dictionary w/o training	51.13	And Dr. Ehud Ur, like some other experts, doubts the possibility of a cure for diabetes, pointing out that these results have no connection to people who are already afflicted with type 1 diabetes.
Selective dictionary w/ training	70.92	Dr. Ehud Ur, like some other experts, is skeptical about the possibility of a cure for diabetes, noting that these findings have no relevance to people with type 1 diabetes.
Selective dictionary w/f training	70.92	Dr. Ehud Ur, like some other experts, is skeptical about the possibility of a cure for diabetes, noting that these findings have no relevance to people with type 1 diabetes.
Selective dictionary w/f/b training	70.85	And Dr. Ehud Ur, like some other experts, is skeptical about the possibility of a cure for diabetes, noting that these findings have no relevance to people with type 1 diabetes.
SeEM w/DSK-BLI	72.89	And Dr. Ehud Ur,like some other experts, who is skeptical about whether of a cure for diabetes, noting that these findings have no relevance to the people afflicted with type 1 diabetes.'
SeEM w/GPT-BLI	74.51	And Dr. Ehud Ur, like some other experts, is skeptical about whether people can recover from diabetes, noting that these findings have no relevance to people with type 1 diabetes.'
SeEM w/DSK-MLI	74.57	Dr. Ehud Ur, like some other skeptical experts, is skeptical about whether a cure exists for diabetes, noting that these findings have no relevance to people who are afflicted with type 1 diabetes.'
SeEM w/GPT-MLI	85.04	Dr. Ehud Ur, like some other experts, is skeptical about whether diabetes can be cured, noting that these findings have no relevance to people who suffer from type 1 diabetes.

Figure 3: Case study on translating following our proposed framework on XE

Source Sentence (Central Kanuri (Latin script)): Karuwu kutta jamibe diye səkə kasadə duwo tədin ma sau kəla lamarə hakku kamubero gaubedəro katap gotəbebe diye ngaworo walətəna kuru kərmaro nasha warakə hawarəbe shima diyen amma million kadda mbeji.		
Target Sentence (Yoruba): Ìkórira agbègbè náà fa agbára láti kọ ilànà tó ní ṣe pẹ̀lú ìbálòpọ́ fún ojú òpó tó gba àlejò tó mílíọ̀nù mídìà gbangba.		
Model Name	Chrf	Translation
Vanilla	20.14	Ìbànújẹ́ tó lágbára yìí ṣẹlẹ̀ nítorí ìṣe aṣejù tí wọ́n ṣe, tí kò tíì mọ́ títí di àkókò pẹ̀lú, ṣùgbọ́n ó yẹ kí a ṣe ìyípadà tó bá yẹ, kó le dáàbò bo ọ̀pọ̀ ènìyàn lọ́dọọdún.
Random	25.84	Karuwu kọ́ agbègbè náà láti kọ ilànà kẹ̀lẹ́ ní ẹ̀tọ́ kamubero gaubedẹro fún gotəbebe náà ngaworo walẹ́tẹna kuru kọ́rmáro nàṣà warakẹ́ hawarẹ́bé shima náà àmọ́ milionu kẹ́ta méjì.
Noun	33.72	Ìkórira tó wà ní agbègbè yìí jẹ́ pé àwọn ènìyàn kọ ilànà kankan, wọ́n sì ìbálòpọ́ pẹ̀lú àwọn àlejò lórí ojú òpó, nítorí náà, ó ṣòro fún àlejò láti mọ ibi tí wọ́n ti ń lọ. Ṣùgbọ́n àwọn mìíràn mẹta ló ti fìyà jẹ àwọn million ènìyàn.
Adjective	21.47	Ìwà ọdaran kan ṣẹlẹ nígbà tí obìnrin àjálù kan pàdé àwọn ọkùnrin méjì lẹ́yìn àkókò tó pé, wọ́n sì ṣe ìmúlòlùfẹ́ pẹ̀lú rẹ̀ lọ́nà tí kò bófin mu ní gbangba, níbi tí ọ̀pọ̀ ènìyàn tó lé ní milionù méjì wà.
Noun and Adjective	38.59	Karuwu jẹ́ agbègbè tí agbára kọ́ sí, ṣugbọn àwọn ẹ̀ka láti ṣe àtìlẹyìn àti ẹ̀tọ́ àwọn olùkọ́ ní agbègbè náà, pẹ̀lú àwọn ohun èlò mídìà gbangba mílíọ̀nù méjì.
Differ in round-trip	29.46	Ìkórira tó lágbára ní agbègbè náà ṣokùnfa ìbànújẹ́ púpọ̀ tí yọrí sí ìlànà tuntun níbi tí wọ́n ṣe pẹ̀lú àkànṣe àtàwọn ilànà tuntun náà, tó jẹ́ pé ó yàtọ̀ gan-an ní àṣà àti ìbáṣepọ̀, tó sì mú kí àwọn ọmọ ilú máa jìyà pupọ̀, ṣùgbọ́n àwọn milionu méjì wà níbẹ̀.
Differ in translation	35.77	Ìkórira tó wà ní agbègbè náà fa agbára búburú méjì tí ó ṣòro láti dá duro, nítorí pé àwọn ẹ̀tọ́ ènìyàn, ìṣedéédé, àti ìbágbépọ̀ ti dàsí. Àlejò kan tó jẹ́ aláìmọ̀ mọ̀ ṣàdédé yáyà wọ̀lé, ó sì dà lórí pé àjálù ń bọ̀, ṣùgbọ́n mílíọ̀nù méjì ló ti ní ipa lórí rẹ̀.
SLoW	35.93	Ìkórira kó gbogbo ẹ̀kọ́ kọ́ lórí bi a ṣe kọ́ láti gba àlejò tó jẹ́ ìbálòpọ́, gbogbo ohun tó wà fún àlejò náà kò rọrùn, ṣùgbọ́n mílíọ̀nù méjì ni.
Selected dictionary w/o training	37.25	Ìkórira tó wà nínú agbègbè yìí fa agbára tó n pọ̀ síi, tí ó sì ṣe pẹ̀lú ìbálòpọ́ àti ìtàn mídìà, ṣùgbọ́n ó ti ní ipa àìmọ̀ye lẹ́yìn tí àwọn ẹ̀rọ ìròyìn àti àtẹ̀jáde yí padà ṣùgbọ́n ẹgbẹ̀rún mẹjọ ni wọn ṣàkóso.
Inverse selected dictionary w/o training	22.25	Karuwu àti kutta jamibe náà səkə kasadə láti kọ ilànà kẹ́ lamarə hakku kamubero gaubedəro, fún gotəbebe náà, ngaworo àlejò tó kərmaro nasha warakə hawarəbe, shima náà ṣùgbọ́n million kadda mbeji.
Selective dictionary w/ training	41.38	Ìkórira kútà agbègbè ń fa agbára dúró, tí ó sì ń sọ̀kalẹ̀ lórí ìbálòpọ́; àwọn míì nípa mídìà gbangba ni wọ́n ń sọ̀rọ̀ nípa rẹ̀, ṣùgbọ́n ó tó mílíọ̀nù méjì.
Selective dictionary w/f training	42.41	Ìkórira tuntun ti farahàn ní agbègbè kan ti ni agbára láti yí ilànà ìbálòpọ́ padà, tí ó sì ti di gbangba lórí mídìà pé ó kan mílíọ̀nù méjì.
Selective dictionary w/f/b training	43.30	Ìkórira pẹ̀lú agbègbè náà ní agbára owó tó pọ̀ láti ṣe lamarè àwọn ìbálòpọ́, ìjọpọ̀ àti àwọn míìdíà gbangba tó jẹ́ mílíọ̀nù méjì.
SeEM w/DSK-BLI	32.26	Ìkórira pọ̀ ní agbègbè yìí, nítorí agbára àwọn olùdarí àti àwọn ilànà tí wọ́n fi ń ṣàkóso, tí wọ́n ṣe ní àfihàn, pẹ̀lú ẹ̀rọ ìbánisọ̀rọ̀ tí ń gbangba gbogbo àkọsílẹ̀, ṣùgbọ́n mẹ̀wàá mílíọ̀nù kì í tó ìkànsí.
SeEM w/GPT-BLI	44.99	Ìkórira ti pọ̀ ní agbègbè náà, wọn fa agbára sí i, wọ́n sì tẹ̀síwájú nípa ilànà tí tó ní ẹ̀tọ́, pẹ̀lú ìbálòpọ́ tó pọ̀. Nínú mídìà, ó gbangba pé ó ju milionu méjì lọ.
SeEM w/DSK-MLI	43.74	Ìkórira pọ̀ ní agbègbè tí agbára ńlá wa, ilànà tó dáa ló wà, pẹ̀lú ìbálòpọ́ àti ìbáṣepọ̀ tó pọ̀, àwọn mídìà gbangba ní ìmọ̀lára rẹ̀, ṣùgbọ́n ẹgbẹ̀rún milionu méjì wà.
SeEM w/GPT-MLI	45.50	Ìkórira agbègbè náà ní agbára, ilànà àti lamarè, àti ìbálòpọ́ tó pọ̀, pẹ̀lú mídìà gbangba tó jẹ́ million méjì.

Figure 4: Case study on translating following our proposed framework on XX

Chrf Util. bam_Latn	Chrf Util. tel_Telu	Chrf Util. mal_Mlym	Chrf Util. mar_Deva	Chrf Util. urd_Arab	Chrf Util. quy_Latn	Chrf Util. khm_Khmr	Chrf Util. kin_Latn	Chrf Util. vec_Latn	Chrf Util. san_Deva
0.6129 0.8287	0.7294 0.5301	0.6051 0.2484	0.7740 0.6843	0.7332 0.7343	0.7392 0.7249	0.5735 0.5782	0.7787 0.6975	0.8382 0.8373	0.7602 0.6610
0.6355 0.8573	0.7671 0.6189	0.7013 0.3633	0.7963 0.7401	0.7458 0.7615	0.7528 0.7682	0.6230 0.6701	0.8034 0.7443	0.8485 0.8723	0.7901 0.7316
0.6388 0.8707	0.7949 0.6615	0.7410 0.4414	0.8077 0.7735	0.7556 0.7906	0.7673 0.7998	0.6318 0.7113	0.8120 0.7719	0.8602 0.8892	0.8004 0.7674
0.6464 0.8857	0.8064 0.7107	0.7633 0.4979	0.8088 0.8028	0.7596 0.8091	0.7709 0.8139	0.6470 0.7419	0.8229 0.8000	0.8647 0.9050	0.8063 0.7989
0.6518 0.9036	0.8143 0.7660	0.7858 0.5708	0.8187 0.8406	0.7572 0.8287	0.7790 0.8347	0.6607 0.7735	0.8427 0.8330	0.8678 0.9180	0.8146 0.8326
kan_Knda	ewe_Latn	bod_Tibt	mos_Latn	cjk_Latn	luo_Latn	dyu_Latn	mag_Deva	tur_Latn	ban_Latn
0.6515 0.3886	0.5606 0.8132	0.6385 0.3890	0.5360 0.8255	0.6613 0.8572	0.5248 0.8038	0.5460 0.8280	0.7367 0.8256	0.8420 0.6344	0.7919 0.8207
0.7348 0.5048	0.5877 0.8429	0.6633 0.6026	0.5445 0.8493	0.6735 0.8887	0.5390 0.8347	0.5546 0.8504	0.7516 0.8661	0.8599 0.6691	0.8097 0.8522
0.7765 0.5673	0.5926 0.8574	0.6855 0.7375	0.5586 0.8617	0.6823 0.9019	0.5556 0.8525	0.5612 0.8607	0.7598 0.8890	0.8633 0.7028	0.8135 0.8745
0.7833 0.6206	0.6027 0.8718	0.6729 0.7917	0.5663 0.8746	0.6893 0.9129	0.5608 0.8642	0.5636 0.8696	0.7752 0.9102	0.8621 0.7307	0.8204 0.8896
0.8041 0.6790	0.6064 0.8852	0.6893 0.8253	0.5692 0.8842	0.6907 0.9259	0.5638 0.8765	0.5709 0.8793	0.7754 0.9324	0.8711 0.7677	0.8276 0.9079
hne_Deva	acq_Arab	uig_Arab	aeb_Arab	min_Latn	bjn_Latn	knc_Latn	mlt_Latn	taq_Latn	fra_Latn
0.7025 0.8250	0.7592 0.7019	0.7176 0.5679	0.7304 0.6971	0.8169 0.8279	0.8099 0.8030	0.5209 0.8058	0.8079 0.7515	0.6244 0.8287	0.8854 0.8314
0.7147 0.8633	0.7707 0.7302	0.7436 0.6375	0.7472 0.7449	0.8277 0.8577	0.8211 0.8206	0.5408 0.8273	0.8179 0.7910	0.6403 0.8548	0.8976 0.8491
0.7203 0.8933	0.7858 0.7547	0.7550 0.6888	0.7575 0.7698	0.8343 0.8703	0.8315 0.8428	0.5460 0.8381	0.8168 0.8158	0.6498 0.8646	0.9067 0.8628
0.7371 0.9125	0.7914 0.7736	0.7682 0.7276	0.7672 0.7855	0.8312 0.8823	0.8378 0.8584	0.5468 0.8487	0.8249 0.8355	0.6595 0.8842	0.9028 0.8745
0.7389 0.9347	0.7986 0.7942	0.7791 0.7810	0.7686 0.8036	0.8453 0.8976	0.8393 0.8687	0.5525 0.8654	0.8270 0.8591	0.6618 0.8985	0.9106 0.8847
jpn_Jpan	sag_Latn	taq_Tfng	arb_Latn	kab_Latn	knc_Latn	kam_Latn	scn_Latn	arz_Arab	gla_Latn
0.7090 0.6914	0.4792 0.8326	0.4133 0.7617	0.5681 0.7050	0.6911 0.7862	0.4398 0.8274	0.6880 0.8165	0.8362 0.8371	0.7413 0.7539	0.7513 0.6374
0.7229 0.7227	0.4896 0.8531	0.4260 0.8124	0.5697 0.7394	0.7159 0.8265	0.4559 0.8514	0.7075 0.8397	0.8527 0.8722	0.7667 0.7858	0.7658 0.6850
0.7398 0.7460	0.4940 0.8606	0.4343 0.8438	0.5741 0.7729	0.7292 0.8491	0.4602 0.8646	0.7146 0.8559	0.8647 0.8914	0.7744 0.8125	0.7798 0.7117
0.7319 0.7645	0.5091 0.8726	0.4393 0.8605	0.5777 0.8060	0.7369 0.8666	0.4654 0.8769	0.7192 0.8680	0.8685 0.9043	0.7738 0.8299	0.7865 0.7336
0.7368 0.7890	0.5039 0.8803	0.4119 0.8765	0.5845 0.8361	0.7383 0.8804	0.4694 0.8882	0.7294 0.8814	0.8747 0.9181	0.7855 0.8458	0.7902 0.7566
lmo_Latn	ast_Latn	grn_Latn	zsm_Latn	lug_Latn	azb_Arab	tso_Latn	azj_Latn	dik_Latn	lus_Latn
0.7395 0.7778	0.8478 0.8090	0.6738 0.7720	0.8897 0.8403	0.7059 0.7524	0.6133 0.6799	0.7526 0.7548	0.7349 0.6856	0.5215 0.8082	0.6237 0.7927
0.7483 0.8073	0.8699 0.8428	0.6892 0.8140	0.8993 0.8557	0.7353 0.8046	0.6316 0.7381	0.7832 0.8094	0.7510 0.7336	0.5361 0.8273	0.6326 0.8214
0.7593 0.8247	0.8754 0.8564	0.6962 0.8368	0.9069 0.8669	0.7497 0.8340	0.6380 0.7654	0.7933 0.8320	0.7557 0.7596	0.5445 0.8388	0.6393 0.8446
0.7630 0.8413	0.8801 0.8711	0.7023 0.8588	0.9126 0.8793	0.7627 0.8631	0.6441 0.7884	0.8073 0.8577	0.7634 0.7879	0.5501 0.8545	0.6499 0.8652
0.7711 0.8590	0.8879 0.8816	0.7144 0.8831	0.9210 0.8954	0.7770 0.8878	0.6548 0.8136	0.8136 0.8798	0.7636 0.8158	0.5588 0.8658	0.6553 0.8805

Figure 5: The figure demonstrating the performance of SeEM for translation. Results are reports on SeEM w/GPT-MLI, with directions in translating from English into other languages. The performance is reported on both chrf and dictionary usage.

Chrf Util. bam_Latn	Chrf Util. tel_Telu	Chrf Util. mal_Mlym	Chrf Util. mar_Deva	Chrf Util. urd_Arab	Chrf Util. quy_Latn	Chrf Util. khm_Khmr	Chrf Util. kin_Latn	Chrf Util. vec_Latn	Chrf Util. san_Deva
0.8238 0.8030	0.7937 0.7204	0.7849 0.6911	0.8190 0.7702	0.8011 0.7278	0.7796 0.7060	0.8140 0.6815	0.7963 0.7398	0.8498 0.8228	0.7530 0.6753
0.8486 0.8478	0.8094 0.7508	0.8078 0.7237	0.8255 0.7983	0.8167 0.7578	0.8026 0.7513	0.8274 0.7115	0.8102 0.7612	0.8668 0.8511	0.7783 0.7141
0.8638 0.8715	0.8209 0.7726	0.8136 0.7415	0.8305 0.8175	0.8264 0.7692	0.8185 0.7870	0.8348 0.7299	0.8253 0.7792	0.8721 0.8649	0.7904 0.7438
0.8674 0.8904	0.8280 0.7839	0.8201 0.7613	0.8468 0.8395	0.8386 0.7810	0.8344 0.8087	0.8461 0.7439	0.8363 0.7971	0.8806 0.8749	0.7980 0.7653
0.8768 0.9106	0.8374 0.8068	0.8379 0.7844	0.8611 0.8570	0.8473 0.8039	0.8440 0.8421	0.8585 0.7620	0.8484 0.8187	0.8861 0.8854	0.8034 0.7928
kan_Knda	ewe_Latn	bod_Tibt	mos_Latn	cjk_Latn	luo_Latn	dyu_Latn	mag_Deva	tur_Latn	ban_Latn
0.7824 0.6905	0.7986 0.7836	0.7207 0.5891	0.8403 0.8114	0.8317 0.8392	0.8615 0.8331	0.8294 0.7781	0.8361 0.7934	0.8068 0.7134	0.8681 0.8227
0.7976 0.7193	0.8216 0.8219	0.7372 0.6543	0.8694 0.8525	0.8550 0.8721	0.8844 0.8597	0.8570 0.8173	0.8532 0.8222	0.8134 0.7359	0.8811 0.8496
0.8050 0.7414	0.8298 0.8402	0.7539 0.7023	0.8782 0.8692	0.8651 0.8966	0.8926 0.8812	0.8639 0.8413	0.8592 0.8337	0.8204 0.7505	0.8930 0.8679
0.8106 0.7568	0.8404 0.8623	0.7661 0.7459	0.8897 0.8909	0.8666 0.9161	0.8934 0.8964	0.8804 0.8633	0.8739 0.8537	0.8285 0.7689	0.9126 0.8855
0.8233 0.7789	0.8471 0.8808	0.7774 0.7658	0.8906 0.9036	0.8822 0.9345	0.8986 0.9103	0.8817 0.8831	0.8790 0.8706	0.8460 0.7927	0.9100 0.9029
hne_Deva	acq_Arab	uig_Arab	aeb_Arab	min_Latn	bjn_Latn	knc_Latn	mlt_Latn	taq_Latn	fra_Latn
0.8199 0.7997	0.8161 0.6932	0.7258 0.6338	0.7976 0.6970	0.8416 0.8055	0.8276 0.7608	0.8340 0.8005	0.9020 0.8376	0.8310 0.8031	0.8331 0.7993
0.8426 0.8340	0.8353 0.7140	0.7462 0.6673	0.8122 0.7171	0.8642 0.8396	0.8431 0.7876	0.8619 0.8329	0.9080 0.8548	0.8511 0.8460	0.8382 0.8121
0.8509 0.8528	0.8436 0.7396	0.7506 0.6968	0.8237 0.7304	0.8749 0.8583	0.8574 0.8183	0.8668 0.8547	0.9129 0.8610	0.8680 0.8657	0.8454 0.8202
0.8588 0.8679	0.8582 0.7542	0.7695 0.7262	0.8302 0.7420	0.8798 0.8742	0.8713 0.8368	0.8713 0.8713	0.9165 0.8692	0.8765 0.8862	0.8527 0.8359
0.8752 0.8860	0.8643 0.7760	0.7832 0.7556	0.8378 0.7663	0.8905 0.8928	0.8741 0.8605	0.8804 0.8831	0.9218 0.8778	0.8792 0.9008	0.8566 0.8467
jpn_Jpan	sag_Latn	taq_Tfng	arb_Latn	kab_Latn	knc_Latn	kam_Latn	scn_Latn	arz_Arab	gla_Latn
0.7645 0.6531	0.8168 0.7883	0.8370 0.7377	0.7590 0.7104	0.8305 0.8069	0.8016 0.7977	0.8319 0.7955	0.8122 0.8002	0.7894 0.7262	0.8065 0.6582
0.7779 0.6852	0.8570 0.8219	0.8516 0.7875	0.7780 0.7410	0.8590 0.8427	0.8243 0.8329	0.8472 0.8287	0.8237 0.8214	0.8016 0.7499	0.8206 0.6817
0.7853 0.7021	0.8689 0.8482	0.8658 0.8186	0.7910 0.7614	0.8682 0.8622	0.8328 0.8571	0.8573 0.8477	0.8335 0.8374	0.8102 0.7663	0.8322 0.6950
0.7964 0.7217	0.8739 0.8646	0.8679 0.8413	0.7937 0.7804	0.8766 0.8784	0.8430 0.8707	0.8691 0.8629	0.8399 0.8473	0.8181 0.7829	0.8344 0.7039
0.8137 0.7454	0.8843 0.8877	0.8740 0.8583	0.8139 0.8032	0.8907 0.8952	0.8497 0.8960	0.8814 0.8818	0.8570 0.8648	0.8266 0.8003	0.8431 0.7179
lmo_Latn	ast_Latn	grn_Latn	zsm_Latn	lug_Latn	azb_Arab	tso_Latn	azj_Latn	dik_Latn	lus_Latn
0.8720 0.7742	0.8538 0.7733	0.8104 0.7665	0.8638 0.8237	0.8336 0.7982	0.7229 0.6318	0.8523 0.7907	0.7213 0.6979	0.8704 0.7988	0.7354 0.7304
0.8833 0.8014	0.8461 0.7922	0.8317 0.8080	0.8757 0.8451	0.8578 0.8305	0.7433 0.6896	0.8666 0.8254	0.7385 0.7314	0.8765 0.8245	0.7626 0.7790
0.8944 0.8128	0.8509 0.8024	0.8398 0.8326	0.8801 0.8605	0.8620 0.8527	0.7536 0.7005	0.8741 0.8388	0.7485 0.7483	0.8898 0.8409	0.7819 0.8092
0.8941 0.8203	0.8539 0.8117	0.8561 0.8586	0.8903 0.8707	0.8778 0.8732	0.7679 0.7238	0.8805 0.8522	0.7611 0.7679	0.8902 0.8536	0.7949 0.8364
0.8975 0.8291	0.8668 0.8276	0.8733 0.8826	0.8979 0.8833	0.8787 0.8921	0.7894 0.7631	0.8910 0.8688	0.7722 0.7932	0.8995 0.8690	0.8108 0.8653

Figure 6: The figure demonstrating the performance of SeEM for translation. Results are reports on SeEM w/GPT-MLI, with directions in translating from other languages into English. The performance is reported on both chrf and dictionary usage.

Chrf Util. knc→yor	Chrf Util. wol→bos	Chrf Util. lus→tzm	Chrf Util. por→dik	Chrf Util. knc→tum	Chrf Util. wol→hne	Chrf Util. gaz→min	Chrf Util. sin→asm	Chrf Util. pap→tir	Chrf Util. mri→ban
0.4342 0.3300	0.5399 0.4403	0.3674 0.2929	0.4497 0.4935	0.4807 0.2913	0.3717 0.3424	0.4386 0.2387	0.5367 0.3464	0.3572 0.4172	0.4941 0.2621
0.4993 0.5191	0.6508 0.5782	0.4565 0.4981	0.4729 0.5713	0.5618 0.4573	0.4796 0.5440	0.5197 0.4192	0.6545 0.5424	0.4699 0.5380	0.5693 0.4237
0.5519 0.6431	0.6744 0.6438	0.5030 0.6302	0.4942 0.6214	0.6066 0.5677	0.4990 0.6764	0.5652 0.6278	0.6847 0.6371	0.4337 0.5903	0.6393 0.5644
0.5631 0.7213	0.7005 0.7226	0.5430 0.7148	0.4870 0.6512	0.6285 0.6641	0.5280 0.7968	0.6046 0.6382	0.7143 0.7115	0.4543 0.6292	0.6715 0.6864
0.5964 0.7930	0.7213 0.8155	0.5548 0.7803	0.4895 0.6765	0.6701 0.7881	0.5729 0.8888	0.6356 0.7437	0.7377 0.7857	0.4569 0.6546	0.7064 0.7973
tir→xho	lua→ron	sag→ceb	tgk→zho	swe→ibo	hin→pap	pbt→sot	tgk→min	awa→sat	mai→ayr
0.4719 0.1224	0.6272 0.5246	0.5744 0.3091	0.4388 0.3259	0.6134 0.3106	0.6357 0.4162	0.5683 0.3933	0.5810 0.4224	0.4468 0.5821	0.4620 0.2540
0.5438 0.2528	0.6633 0.6243	0.6292 0.4496	0.4925 0.4706	0.6626 0.4490	0.6691 0.5479	0.6585 0.6391	0.6638 0.6020	0.5087 0.7623	0.5437 0.4272
0.5979 0.3809	0.6752 0.6907	0.6533 0.5357	0.5283 0.5709	0.6855 0.5226	0.6802 0.6360	0.6814 0.7391	0.6945 0.6791	0.5329 0.8392	0.5979 0.5689
0.6566 0.5174	0.6944 0.7469	0.6785 0.6233	0.5443 0.6507	0.7075 0.5789	0.7196 0.7141	0.7111 0.8128	0.7252 0.7439	0.5434 0.8748	0.6393 0.6735
0.6902 0.6568	0.6981 0.8039	0.6965 0.7024	0.5745 0.7242	0.7125 0.6318	0.7190 0.7993	0.7164 0.8772	0.7443 0.8080	0.5497 0.9006	0.6934 0.7920
plt→mar	bjn→ces	yor→ceb	nno→ssw	bug→mai	ltg→scn	por→ace	mai→tso	lij→ace	cus→slv
0.5440 0.3543	0.5485 0.3429	0.5546 0.3461	0.5429 0.2273	0.4116 0.3842	0.5163 0.3133	0.5665 0.1957	0.6892 0.3852	0.4109 0.3192	0.5681 0.3501
0.6145 0.5134	0.6069 0.4727	0.6124 0.4947	0.6225 0.3811	0.4990 0.5749	0.5770 0.4517	0.6221 0.6567	0.7062 0.5343	0.5620 0.3506	0.6525 0.5103
0.6771 0.6305	0.6275 0.5451	0.6401 0.5897	0.6590 0.4888	0.5363 0.7011	0.6033 0.5427	0.6471 0.7392	0.7438 0.6300	0.6144 0.6837	0.6861 0.6060
0.7334 0.7278	0.6602 0.6261	0.6600 0.6682	0.6910 0.5898	0.5669 0.7926	0.6278 0.6294	0.6695 0.7995	0.7593 0.7071	0.6621 0.7810	0.7124 0.6969
0.7504 0.8130	0.6951 0.6967	0.6971 0.7578	0.7429 0.7370	0.5897 0.8622	0.6548 0.7263	0.6941 0.8511	0.7789 0.8021	0.6733 0.8404	0.7303 0.7927
bug→lij	mri→bjn	dyu→npl	plt→hat	nso→nya	kmr→ory	hat→hin	xho→tsn	tso→kmr	kon→dan
0.5944 0.3336	0.4185 0.1272	0.5858 0.4656	0.6198 0.3906	0.5795 0.3647	0.4159 0.1133	0.5572 0.5173	0.5339 0.3918	0.5807 0.3853	0.6351 0.5240
0.6322 0.4756	0.4614 0.2633	0.6190 0.6089	0.6589 0.5057	0.6445 0.5053	0.5012 0.2538	0.6196 0.6414	0.5845 0.5273	0.6319 0.5427	0.6862 0.6776
0.6550 0.5765	0.5080 0.4102	0.6649 0.6896	0.6731 0.5949	0.6843 0.6303	0.5941 0.4079	0.6541 0.7226	0.6172 0.6205	0.6687 0.6528	0.7156 0.7558
0.6734 0.6623	0.5379 0.5316	0.6980 0.7538	0.6939 0.6693	0.7215 0.7095	0.6651 0.5542	0.6790 0.7921	0.6375 0.7148	0.6920 0.7526	0.7240 0.8146
0.6924 0.7541	0.6124 0.6374	0.7014 0.8028	0.7249 0.7479	0.7596 0.8150	0.7174 0.6822	0.6921 0.8511	0.6609 0.7766	0.7125 0.8282	0.7452 0.8845
ace→ltg	kon→hin	hne→kmr	por→min	plt→swe	npi→hne	tso→ace	swe→ayr	tso→ssw	ace→aws
0.4371 0.2577	0.5638 0.5760	0.6853 0.2487	0.7079 0.5989	0.7738 0.5223	0.5852 0.3114	0.4785 0.2783	0.5915 0.4398	0.5292 0.1821	0.4763 0.2721
0.5261 0.4396	0.6319 0.7327	0.7393 0.4293	0.7516 0.6967	0.8198 0.6462	0.6551 0.4763	0.5752 0.4656	0.6535 0.5909	0.5842 0.3032	0.5574 0.4644
0.5888 0.5588	0.6609 0.8006	0.7715 0.5539	0.7758 0.7453	0.8527 0.7163	0.7005 0.6091	0.6230 0.5983	0.6888 0.6860	0.6214 0.3970	0.5955 0.6179
0.6345 0.6710	0.6648 0.8566	0.8145 0.6841	0.7815 0.7782	0.8651 0.7738	0.7463 0.7101	0.6602 0.7155	0.7230 0.7488	0.6666 0.5054	0.6443 0.7299
0.6749 0.7773	0.6817 0.9094	0.8558 0.7998	0.7909 0.8202	0.8945 0.8488	0.7857 0.8216	0.7130 0.8226	0.7403 0.8215	0.7242 0.6638	0.6603 0.8191

Figure 7: The figure demonstrating the performance of SeEM for translation. Results are reports on SeEM w/GPT-MLI, with directions in translating from other languages into other languages. The performance is reported on both chrf and dictionary usage.

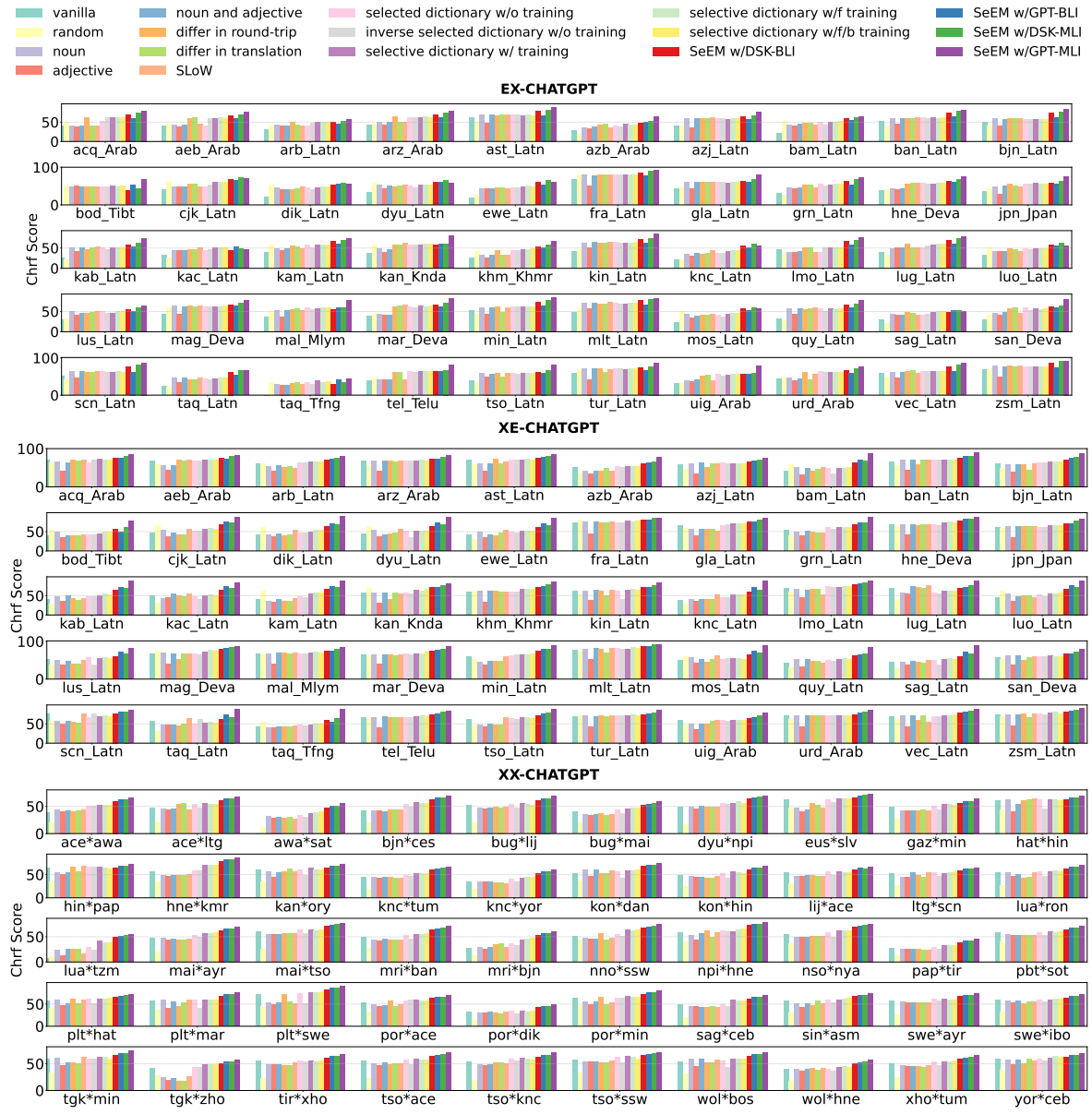


Figure 8: The figure presents the translation performance of SeEM, evaluated using GPT-4o-mini. Results are reported for three translation directions: (1) English to other languages, (2) other languages to English, and (3) between non-English languages. Performance is measured using chrF++ scores.

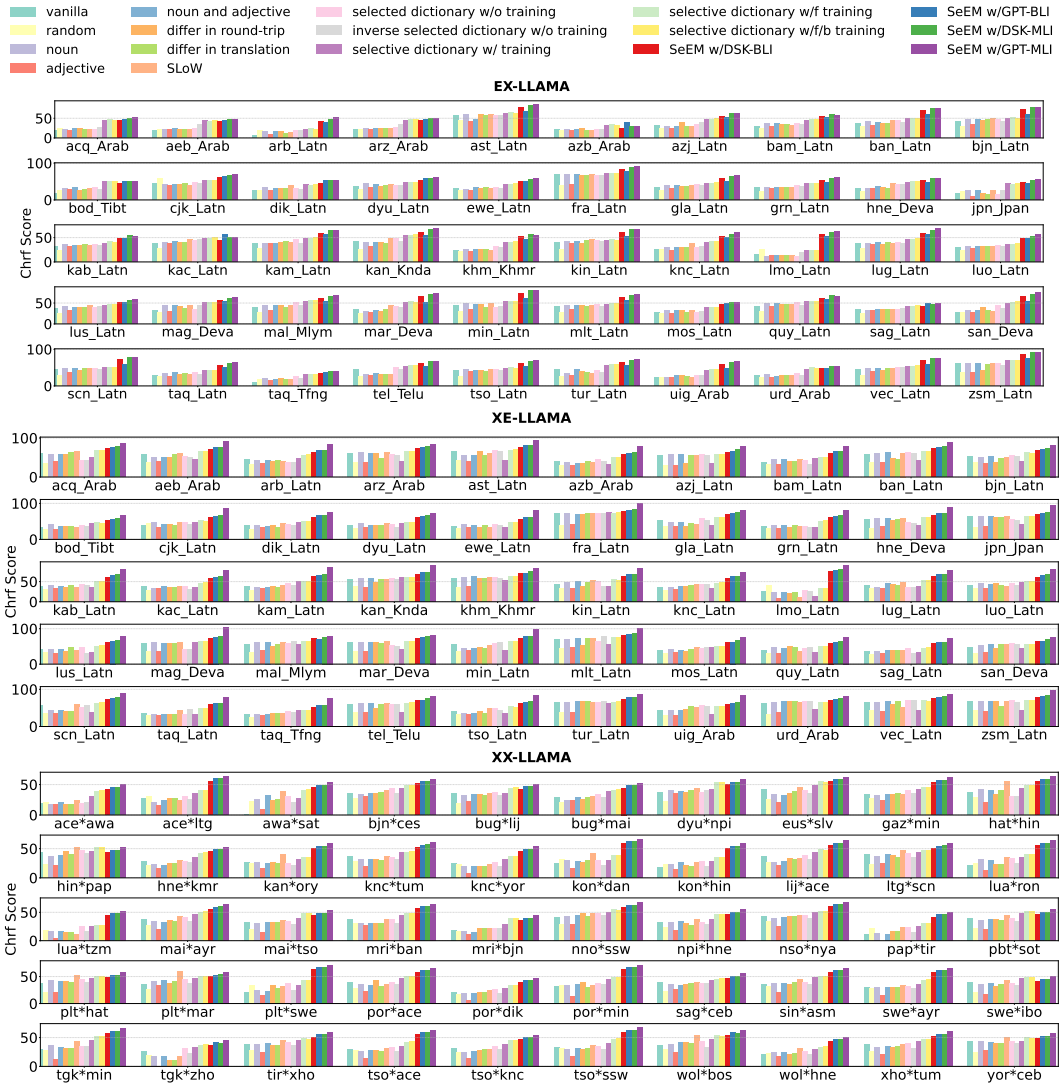


Figure 9: The figure presents the translation performance of SeEM, evaluated using LLaMA3. Results are reported for three translation directions: (1) English to other languages, (2) other languages to English, and (3) between non-English languages. Performance is measured using chrF++ scores.

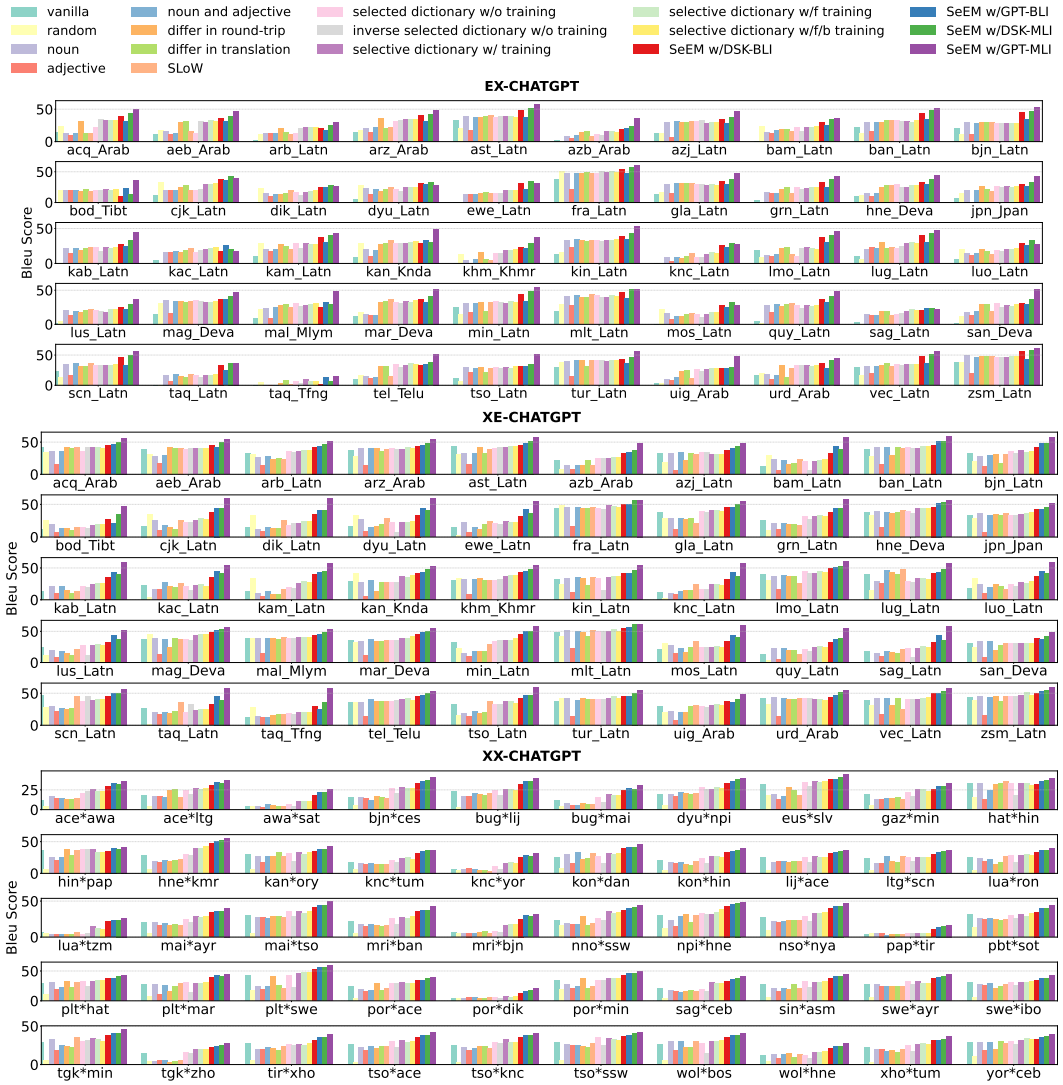


Figure 10: The figure demonstrating the performance of SeEM for translation. Results are reports on GPT-4o-mini, with directions in translating from English into other languages, from other languages into other languages, and from other languages into other languages. The performance is reported in BLEU scores.



Figure 11: The figure demonstrating the performance of SeEM for translation. Results are reports on Llama, with directions in translating from English into other languages, from other languages into other languages, and from other languages into other languages. The performance is reported in BLEU scores.