SCALE: Synergized Collaboration of Asymmetric Language Translation Engines

Anonymous ACL submission

Abstract

In this paper, we introduce SCALE, a collaborative framework that connects a compact Specialized Translation Model (STM) and a general-purpose Large Language Model (LLM) as one unified translation engine. By introducing translation from STM into the triplet in-context demonstrations, SCALE unlocks refinement and pivoting ability of LLM, thus 1) mitigating language bias of LLMs and parallel data bias of STMs, 2) enhancing LLM speciality without sacrificing generality, and 3) facilitating continual learning in a LLM-tuning-free way. Our comprehensive experiments show that SCALE significantly outperforms both LLMs (GPT-4, GPT-3.5) and supervised models (NLLB, M2M) in either high-resource or challenging low-resource settings. Moreover SCALE demonstrates the scalability by only updating the lightweight STM and witness consistent system improvement, an averaged 4 BLEURT score across 4 languages without tuning LLM. Interestingly, SCALE could also effectively exploit the existing language bias of LLMs by using an English-centric STM as a pivot to conduct translation between any language pairs, outperforming GPT-4 by an average of 6 COMET points across eight translation directions. Furthermore we provide an in-depth analysis of SCALE's robustness, translation characteristics, latency costs and inherent language bias, providing solid foundation for future studies exploring the potential synergy between LLMs and more specialized models.

1 Introduction

Large Language Models (LLMs) have recently revolutionized the field of natural language processing (OpenAI, 2023; Touvron et al., 2023a; Anil et al., 2023; Peng et al., 2023) and brought a paradigm shift in machine translation (MT) by delivering exceptional performance without relying on bilingual data (Brown et al., 2020; Garcia et al., 2023). Moreover, as a unified multi-task learner, LLMs represent a substantial step towards artificial general intelligence (Bubeck et al., 2023), with the potential to transcend not only the language barriers emphasized by previous MT research but also cultural boundaries (Yao et al., 2023).

Despite their advancements, LLM-based translation systems still confront several challenges. Firstly, there exists a significant language bias towards English (e.g., 92.1% of the GPT-3 pretraining corpus is English, while French, the second largest, represents only 1.8%), which significantly constraints multilingual ability, especially for those low-resource languages (Scao et al., 2022; Hendy et al., 2023; Zhang et al., 2023). Secondly, as the go-to approach for improved system performance, fine-tuning is non-trivial for LLMs due to (1) the trade-off between speciality and generality (Lin et al., 2023; Cheng et al., 2023a), and (2) the prohibitively high cost associated with tuning large-scale models. In contrast, traditional Specialized Translation Models (STMs)—those based on encoder-decoder architecture, trained with labeled data and significantly smaller in size (Sutskever et al., 2014; Vaswani et al., 2017)—serve as specialists for specific translation tasks and could be efficiently fine-tuned. Nevertheless, due to restricted model capacity, these models exhibit limitations in general language capabilities and may be prone to parallel data bias, such as the memorization of low-quality samples (Raunak et al., 2022).

In this paper, we demonstrate for the first time the possibility to unify these two asymmetric translation engines in a single framework. Our work, SCALE, connects LLMs and STMS by utilizing the LLM's most enigmatic capability: in-context learning. Rather than employing source-target pairs as in conventional few-shot translation (Garcia et al., 2023; Vilar et al., 2023), SCALE would first sample translations from a STM and then use triplets as in-context demonstrations consisting of a source sentence, an STM-generated set and a target sentence,

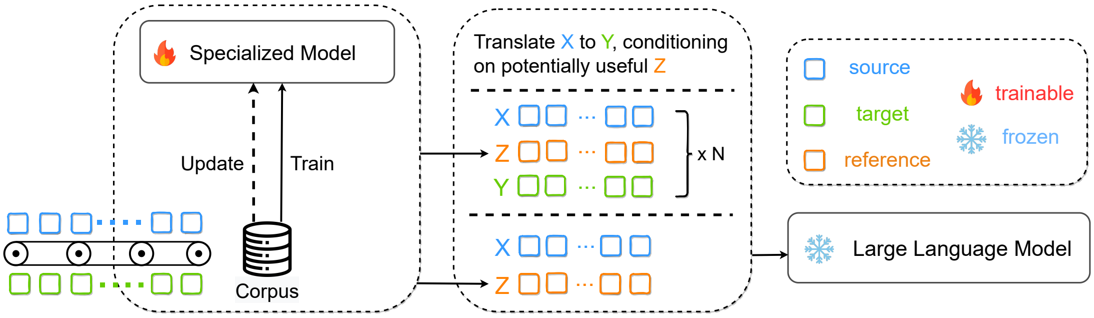

Figure 1: The SCALE framework, comprised of a lightweight specialized model and a frozen large language model with triplet in-context demonstrations.

which unlocks refinement and pivoting ability of LLMs. With SCALE, we could (1) mitigate both language bias of LLMs by utilizing an STM that concentrates on a specific language pair, and parallel data bias of STMs by using a general-purpose LLM as the main body of the system; (2) enhance the speciality of LLMs without compromising generality; (3) facilitate continual learning within the framework by updating only the lightweight STM, thus avoiding expensive LLM fine-tuning. By employing SCALE, we create a more efficient and effective system that combines the best of both translation engines.

Our comprehensive experiments reveal that SCALE considerably outperforms LLMs (e.g., GPT-4) and STMs (e.g., NLLB) even in the challenging low-resource setting. Moreover, in Xhosa→English direction, SCALE experiences consistent improvement by a 4 BLEURT score without tuning LLM and surpasses few-shot GPT-4 by 2.5 COMET score and 3.8 BLEURT score when equipped with a compact model consisting of merely 600M parameters. Interestingly, SCALE could exploit the existing language bias of LLMs by using an English-centric STM as a pivot to conduct translation between any language pairs, outperforming GPT-4 by an average of 6 COMET points across eight translation directions. Furthermore, we conduct an in-depth analysis of the robustness, translation characteristics, latency costs and inherent language bias associated with SCALE. Our findings provide valuable insights and encourage further research in the synergy between LLMs and small specialized models.

2 The SCALE Framework

In this section, we present SCALE and provide an overview illustrated in Figure 1. Popularized by GPT-3 (Brown et al., 2020), In-context Learning (ICL) allows LLMs to perform a wide variety of tasks, even newly created ones (Bills et al., 2023), by providing a few demonstrations. For a translation task from a source language $\mathcal{X}$ to a target language $\mathcal{Y}$, an LLM with parameters θ carries out ICL by conditioning on k source-target paired examples $\mathbb{E} = (x_1, y_1) \oplus (x_2, y_2) \oplus ... (x_k, y_k)$ and the test source sentence x, generating the target y in an auto-regressive manner as $y_t \sim p_\theta(y_t|\mathbb{E}, x, y_{<t})$. In this scenario, the LLM must analyze the provided examples to discern the input distribution, output distribution, input-output mapping, and formatting to successfully complete the task (Press et al., 2022; Wei et al., 2023). Different from conventional ICL, SCALE introduces an intermediate variable $\mathbb{Z}$ as reference between source x and target y, transforming each demonstration example into a triplet $(x, \mathbb{Z}, y)$. The variable $\mathbb{Z}$ is a generation set sampled from a specialized translation model $\mathbf{M}_{\mathcal{X} \mapsto \mathcal{Y}}$ trained on a labeled dataset. The final input to the LLM consists of the instruction, demonstrations, and source sentence combined in a prompt template: $\mathcal{T}((x_1, \mathbb{Z}_1, y_1) \oplus (x_2, \mathbb{Z}_2, y_2) ... \oplus (x_k, \mathbb{Z}_k, y_k)), (x, \mathbb{Z}))$. Unlike language understanding tasks that have fixed label set (Xu et al., 2023), the hypothesis space of translation model is actually infinite, so we could sample multiple generation paths from STM for one single source sentence to provide a more comprehensive generation guide for LLM. The SCALE framework, though conceptually straightforward, demonstrates several advantages over STMs and LLMs:

Refinement For $\mathcal{X}$ to $\mathcal{Y}$ translation task, when the intermediate variable $\mathbb{Z}$ is from $\mathbf{M}_{\mathcal{X} \mapsto \mathcal{Y}}(x)$, SCALE essentially conduct few-shot learning in a multi-task way by introducing an additional refinement task. Refinement has long been proved effective in MT (Xia et al., 2017; Cheng et al., 2022). In this refinement process, we pass sampled

sentences and their confidence score (token level probability) from STM to an LLM. The LLM then digests the information carried by the sampled set and infers the generation space of the STM, which guides the LLM to generate the output that is more consistent with the local data distribution (Xu et al., 2023). And since the final translation is delivered by an LLM, SCALE could also mitigate the parallel data bias from STMs and exhibit robustness by not merely copying and pasting the draft translation from STMs as shown in §5.3.

Pivoting Considering the predominantly English-centric nature of most LLMs (Brown et al., 2020; Touvron et al., 2023a), SCALE could employ an intermediate variable $\mathbb{Z}$ from $\mathbf{M}_{\mathcal{X}\mapsto\text{English}}(x)$ where the target language $\mathcal{Y}$ is not necessarily English. And here $\mathbb{Z}$ serves as a pivot point for LLMs to enhance their understanding of the source sentence and yield improved translations. This can also be regarded as a form of knowledge transfer from high-resource languages to low-resource languages (Kim et al., 2019).

Scalability A significant limitation of the existing LLM-based translation systems is the inherent complexity of LLM continual learning. This complexity arises from several factors, including the balance between speciality and generality (Lin et al., 2023), the catastrophic forgetting problem (Yong et al., 2023), and the substantial computational demands. In contrast, SCALE offers a more efficient and streamlined approach for scalability. By exclusively and effectively updating the lightweight $\mathbf{M}_{\mathcal{X}\mapsto\cdot}$ component, SCALE ensures that the LLM remains untouched, thus preserving its general language capabilities. This selective updating process not only mitigates the issue of catastrophic forgetting but also reduces the computational burden of fine-tuning associated with large-scale models.

3 Experimental Setup

3.1 Dataset

Our evaluation datasets encompass a diverse set of languages, spanning both low- and high-resource ones. To facilitate reproducibility, all our evaluation datasets come from publicly available `devtest` split of Flores-200 (NLLB Team et al., 2022).

3.2 Translation Systems

We compare our approach with cutting-edge academic systems including both specialized models and LLMs, as well as one commercial system, Microsoft Translator. To our knowledge, these models are among the strongest and most representative of their respective categories. For supervised models, we consider **M2M100** (Fan et al., 2021), the first multilingual encoder-decoder translation model that can translate between any pair of 100 languages without relying on English data; **NLLB** (NLLB Team et al., 2022), a supervised translation model suite capable of delivering high-quality translations directly between 200 languages and remains state-of-the-art performance. For LLMs, we consider: **XGLM-7.5B** (Lin et al., 2022), a multilingual generative language models; **GPT-3.5**, a GPT model specially optimized for conversational purpose; **GPT-4** (OpenAI, 2023), the latest version of GPT-series.

We use both GPT-3.5 and GPT-4 from Microsoft Azure OpenAI Service. Without further notice, the number of few-shot samples in LLM and SCALE are set to 10 and the sample selection strategy follows Agrawal et al. (2022). The prompt we use could be found in the Appendix A.

3.3 Evaluation Metrics

Because neural metrics have shown higher correlation with human preference (Freitag et al., 2022; Rei et al., 2020) and are widely adopted by recent literatures (Hendy et al., 2023; Garcia et al., 2023), we mainly evaluate our system with (1) **COMET-22**, a reference-based neural metric (Rei et al., 2022a) combining direct assessments, sentence-level score, and word-level tags from multidimensional quality metrics error annotations, (2) **COMETKiwi**, a refrence-free quality estimation model from Rei et al. (2022b), and (3) **BLEURT** (Sellam et al., 2020), a learnable evaluation metric with a regression model trained on ratings data. For completeness, we also include the results of lexical metrics such as spBLEU (NLLB Team et al., 2022) and chrF++ (Popovic, 2017).

4 Experimental Results

In this section, we conduct various experiments to show the effectiveness of our framework. In §4.1, we verify the effectiveness of the SCALE refinement ability. In §4.2, we focus on non-English pairs to test the pivoting ability of SCALE. In §4.3, we show the scalability of our framework with a fixed LLM and an evolving STM.

	COMET-22	COMETKiwi	BLEURT	COMET-22	COMETKiwi	BLEURT
	eng_Latn→khm_Khmr			khm_Khmr→eng_Latn		
NLLB	76.3	77.8	59.5	86.1	85.4	72.2
M2M100	59.6	58.5	34.2	69.6	71.6	54.0
Microsoft	70.1	70.9	54.7	80.2	80.5	63.3
XGLM	28.1	32.2	19.7	48.6	53.7	21.6
GPT-3.5	68.8	69.3	55.7	73.3	73.0	53.2
GPT-4	74.3	74.7	53.7	84.6	84.0	69.9
SCALE	**79.6**	**80.3**	**61.1**	**87.1**	**85.9**	**73.9**
	eng_Latn→amh_Ethi			amh_Ethi→eng_Latn		
NLLB	84.4	80.7	72.8	86.9	84.5	73.6
M2M100	69.9	68.5	60.7	72.3	72.0	54.8
Microsoft	84.1	80.1	72.6	87.5	84.6	74.7
XGLM	28.0	28.2	20.7	50.2	43.9	17.8
GPT-3.5	66.5	63.2	54.9	58.8	54.2	31.7
GPT-4	77.1	73.4	61.5	83.2	81.9	67.3
SCALE	**84.7**	**81.7**	**74.4**	**88.0**	**85.3**	**75.7**

Table 1: Results of low-resource languages. The best results are in **bold** and the second best are with underscore.

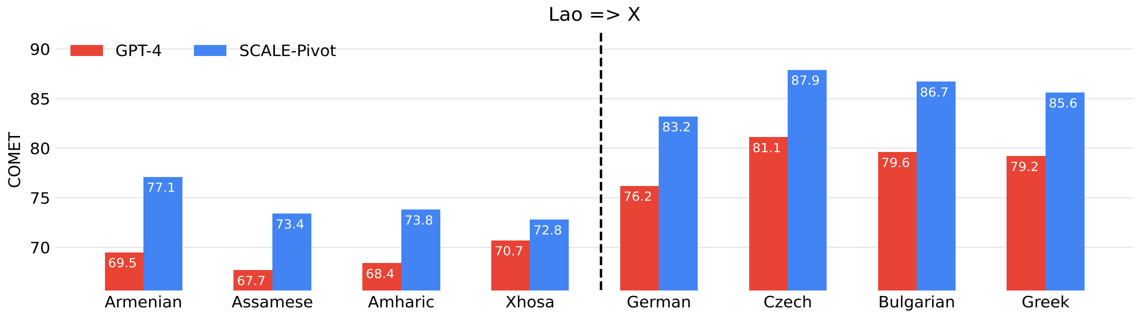

Figure 2: Translation results from Lao to both low- and high-resource languages, where GPT-4 uses few-shot prompting and SCALE-pivot uses English as the pivot language. For more results please refer to Appendix D

4.1 SCALE Refinement

Although LLMs perform comparably with supervised models on high-resource languages, they still struggle with languages with very limited resource (Garcia et al., 2023; Vilar et al., 2023). To validate the generality of our framework, we evaluate in both low- and high-resource setting. For high-resource language, we include English (eng_Latn), Czech (ces_Latn) and Chinese (zho_Hans); for low-resource ones, we include Khmer (khm_Khmr) and Nepali (npi_Deva). We adopt three kinds of baselines as described in §3.2. For supervised NLLB model suite, we choose the NLLB-3.3B version, and for SCALE-refine, the LLM is GPT-4 and the STM is NLLB-3.3B for fair comparison.

The low-resource results are displayed in Table 1. As observed, few-shot LLMs, including GPT-4, significantly trail behind specialized models in all translation directions. In contrast, our framework, by combining LLMs and STMs, demonstrates superior performance over few-shot GPT-4 by an absolute 5 COMET score on average, and surpasses the strong NLLB model in 4/4 directions. The high-resource results are shown in Table 2, leading to the following observation: (1) different from low-resource ones, the few-shot GPT-4 already surpasses supervised counterparts (NLLB and M2M100) by a significant margin; (2) SCALE continues to offer improvements, albeit less substantial than those observed in low-resource settings; (3) SCALE exhibits strong robustness when paired with a less performant STM (especially in eng_Latn→zho_Hans direction).

4.2 SCALE Pivoting

In this section, we demonstrate the performance of SCALE-pivot, in which the variable $\mathbb{Z}$ is not directly pertinent to the current translation directions but functions as a pivot. Specifically, we examine the performance of few-shot GPT-4 and SCALE-pivot on Lao→ $\mathbb{Y}$ translations, where $\mathbb{Y}$ represents

	COMET-22	COMETKiwi	BLEURT	COMET-22	COMETKiwi	BLEURT
	eng_Latn→ces_Latn			ces_Latn→eng_Latn		
NLLB	90.1	84.8	80.3	88.4	85.5	78.6
M2M100	88.2	83.2	77.3	87.3	84.6	76.6
Microsoft	90.3	84.9	79.9	88.5	84.9	77.5
XGLM	27.7	81.9	14.5	57.5	51.6	40.7
GPT-4	92.0	86.8	82.7	**89.4**	85.0	80.3
SCALE	**92.4**	**87.1**	**83.4**	89.2	**86.0**	**80.5**
	eng_Latn→zho_Hans			zho_Hans→eng_Latn		
NLLB	78.0	70.9	58.1	86.1	83.7	74.5
M2M100	83.4	80.8	67.3	85.0	82.9	72.3
Microsoft	86.6	82.1	69.9	86.3	82.9	75.1
XGLM	80.0	75.4	62.0	43.6	74.5	57.0
GPT-4	88.8	**84.7**	73.4	88.0	84.8	77.8
SCALE	**89.1**	**84.7**	**73.6**	**88.3**	**84.9**	**77.9**

Table 2: Results of high-resource languages. The best results are in **bold** and the second best are with underscore.

a language set encompassing both low-resource and high-resource languages. For the low-resource languages, we include Assamese (asm_Beng), Armenian (hye_Armn), Amharic (amh_Ethi), Xhosa (xho_Latn), and we have German (deu_Latn), Czech (ces_Latn), Bulgarian (bul_Cyrl) and Greek (ell_Grek) for the high-resource setting.

The results are presented in Figure 2. Upon examining the GPT-4 performance in isolation, it is evident that the inherent language bias has a considerable impact on translation performance. In particular, the GPT-4 model generally performs well in high-resource settings; however, it tends to struggle in low-resource scenarios. Moreover, our findings highlight that employing SCALE-pivot can effectively enhance the performance of GPT-4 across both low- and high-resource settings. Interestingly, the performance gains achieved through SCALE-pivot are more pronounced in high-resource settings, with an average improvement of 6.8 COMET-22 score compared to 5.2 for low-resource settings. This outcome suggests that incorporating SCALE-pivot can further boost the already strong performance of GPT-4 in high-resource situations, while also providing a notable improvement in low-resource contexts.

4.3 SCALE Scalability

In this section, we explore the scalability of our framework by keeping the LLM fixed and solely updating the STM. Specifically, we use M2M100-12B and NLLB model suite ranging from 600M to 3.3B as our evolving STM. We conduct experiments on the Xhosa → English direction and adopt the prompt format of SCALE-refine. The experimental results are displayed in Figure 3, leading to the following observations: (1) The overall framework can be consistently improved with a fixed LLM and a continuously evolving STM; (2) SCALE, when equipped with a small model containing only 600M parameters, can outperform GPT-4 with an absolute 2.5 COMET-22 score and a 3.8 BLEURT score; (3) Equipped with an STM (M2M100) of relatively lower performance than original few-shot GPT-4, SCALE demonstrates strong robustness by not merely copying and pasting the less satisfactory reference answer provided by M2M100, which we detailedly investigated in §5.3.

Interestingly, we also observe that the growth patterns exhibited by lexical metrics and neural semantic metrics differ. For M2M100 and NLLB-600M as STM, both metrics experience substantial improvement, while for NLLB-1.3B and 3.3B as STM, SCALE maintains the same lexical accuracy while continually enhancing translation performance as measured by neural semantic metrics.

5 Further Analysis

5.1 Translation Characteristics

To gain a deeper understanding of the translation characteristics of different systems (LLMs, STMs, and SCALE) beyond overall translation quality, we employ the following measurements, as suggested by Hendy et al. (2023):

1. **Translation Fluency:** Since LLMs are optimized by predicting the next token, their translations tend to display a language modeling bias that favors fluency over adequacy. To investigate this, we utilize an independently

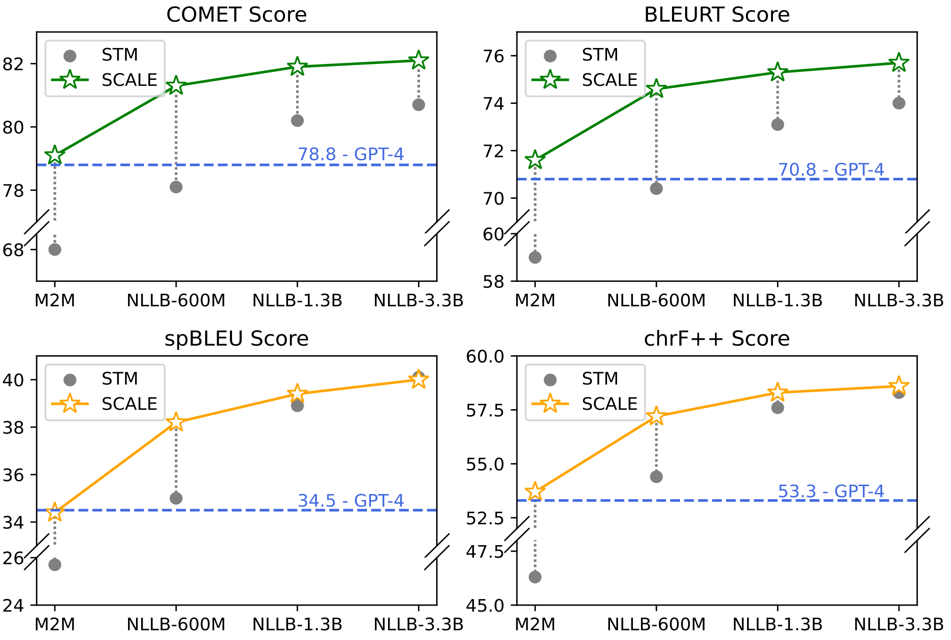

Figure 3: Translation results from Xhosa→English with evolving STMs. More results are in Appendix E

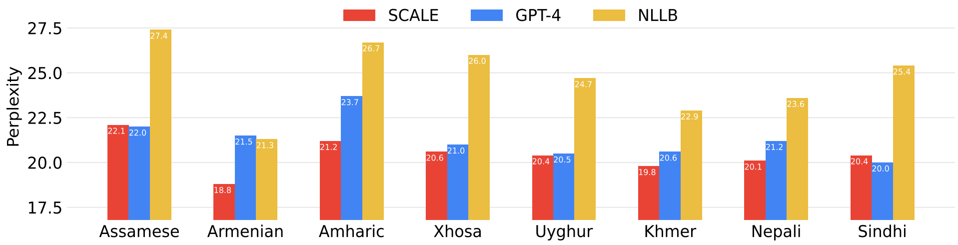

Figure 4: Perplexity score from $\mathbb{X}$→English translation.

trained open-source language model (Llama-2-13B (Touvron et al., 2023b)) to measure the perplexity score of the translation output.

2. **Translation Non-Monotonicity:** This metric evaluates the extent to which a translation adheres to the source sentence's structure, calculating the deviation from the diagonal in the word-to-word alignment. Translations that are more paraphrastic or less literal tend to deviate from closely tracking the source word order across language pairs (Hendy et al., 2023). We apply the non-monotonicity metric proposed by Schioppa et al. (2021).

3. **Unaligned Source Words:** Another measure of literalness is the count of unaligned source words (Hendy et al., 2023). When accounting for quality, less literal translations are likely to include more words that do not align with those in the source sentence.

We present the **Translation Fluency** results of $\mathbb{X}$ → English translation in Figure 4, where $\mathbb{X}$ remains the same as used in Section 4.1. It is evident that regardless of the translation quality delivered by the LLM, whether superior (SCALE) or inferior (GPT-4) compared to the STM (NLLB), the LLM translation generally demonstrates higher fluency than the STM. Additionally, in 6 out of the 8 languages examined, SCALE produces lower perplexity scores than the original GPT-4 output. This suggests that the STM-generated variable $\mathbb{Z}$ can effectively aid the GPT-4 model in further decreasing its generation uncertainty.

For **Non-Monotonicity** and **Unaligned Source Words**, we choose Xhosa→English translation with different STMs, and the results are shown in Figure 5. We also include PPL score for completeness. We find that both the USW and NM scores for STM are higher than those of GPT-4. This indicates that even though STM provides higher

# Path	COMET-22	BLEURT	spBLEU
1	80.4	73.2	35.6
2	81.2	74.3	37.1
3	81.4	74.7	38.0
4	**81.5**	74.8	38.3
5	81.4	**74.9**	**38.4**

Table 3: Translation results from Xhosa→English with multi-path sampling. All the experiments are conducted by one-shot SCALE-refine and only differ in the number of sampled paths from STM.

translation quality, it results in less literal translations. However, for SCALE, it effectively reduces GPT-4's NM score while maintaining a moderate USW score. This suggests that during the SCALE refinement process, the model primarily adheres to the original LLM output structure while taking cues from STM's word selection. We also show several concrete cases in Appendix C.

5.2 Multipath Sampling

In this section, We list the results of multi-path sampling strategy in Table 3. We test with Xhosa→English with one-shot SCALE-refine. The results show that without increasing the shot number in the few-shot learning, using STM to generate more generation paths could consistently improve the overall performance, which could be useful in the extremely low-resource setting where demonstration samples are hard to acquire.

5.3 Ablation

In this section, we conduct an ablation study for each key design in our framework. We examine the following variants: (1) without confidence: This model follows the same setting as the SCALE-refine in §4.1, except that we do not pass the confidence score of each token as input. (2) zero-shot: This variant removes all in-context-learning examples, keeping only the translation instruction and the reference answer from STM. (3) one-shot: This model utilizes only one-shot, in contrast to the ten-shot results presented in §4.1. (4) zero-shot-M2M: This model also implements zero-shot, but the STM used is M2M100, a less performant model than the original few-shot GPT-4. This is employed to assess the robustness of our framework.

The outcomes of our ablation study are showcased in Table 4. It is evident that each component in our framework perform effectively, with the in-context-learning setting providing the most performance gain. This indicates that simply offering a reference answer to the LLM without in-context samples does not adequately guide the model in utilizing those references effectively. Furthermore, the number of ICL examples is also an essential factor in the process.

Regarding the SCALE zero-shot-M2M variant, its performance is significantly inferior to that of the few-shot LLM due to the poor quality of the M2M100 output. From this observation, we can conclude that the robustness of SCALE, as illustrated in Figure 3, primarily stems from the power of in-context learning. This learning approach informs the LLM about which elements to trust and which to disregard, ultimately improving the overall translation performance and robustness.

	COMET-22	BLEURT
M2M100	68.0	59.0
NLLB	80.7	74.0
GPT-4	78.8	70.8
SCALE	**82.1**	**75.7**
w/o confidence	81.6	74.9
zero-shot	81.4	74.8
one-shot	81.7	75.3
zero-shot-M2M	76.4	68.2

Table 4: Ablation results for SCALE.

5.4 Generation Latency

	LLM		SCALE			
# shot	# len.	total	# len.	STM	LLM	total
0	101.37	7.19	161.13	1.87	7.43	9.3
1	198.00	7.46	516.92	1.87	8.33	10.2
10	951.91	9.52	2489.72	1.87	14.17	16.04

Table 5: Latency of LLM (BLOOM-175B) and SCALE (BLOOM-175B + NLLB-3.3B) measured in seconds.

In this section, we conduct a detailed evaluation of the overhead introduced by SCALE in comparison to few-shot LLM. The additional latency arises from two factors: first, the time required to generate the variable $\mathbb{Z}$ for the current source sentence x using STM, and second, the increased latency caused by the LLM due to the extended context. We utilize one of the largest open-source LLMs (BLOOM-176B) for this analysis. As shown in Table 5, we observe that the incurred latency can be primarily attributed to the extended context window due to the quadratic time complexity of the transformer. Exploring methods to accelerate this

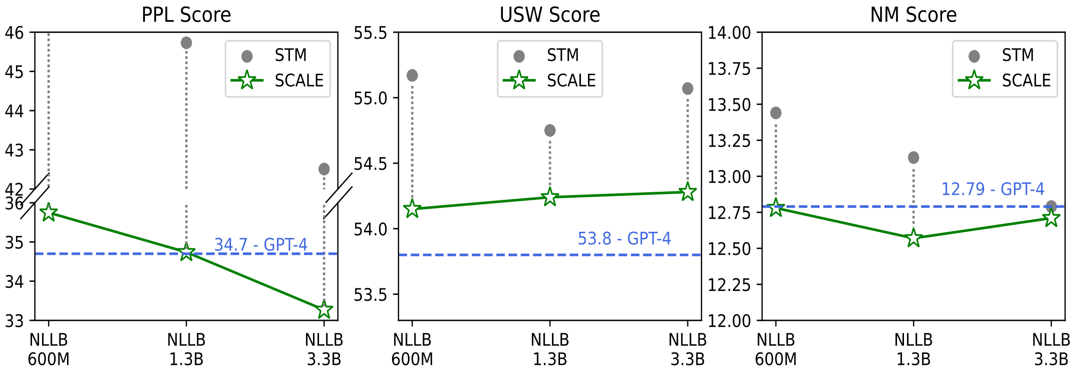

Figure 5: Perplexity, Unaligned Source Words percentage and Non-Monotonicity score for Xhosa→English.

	Javanese	Tamil	Urdu	Amharic
# Speakers	98 M	84.12 M	71.29 M	25 M
GPT-4	83.9/75.8	83.5/80.8	80.0/80.4	77.1/73.4
NLLB	86.4/76.4	86.5/82.9	80.7/80.4	84.4/80.7
SCALE	**86.6/77.5**	**87.8/84.7**	**82.0/81.7**	**84.7/81.7**

Table 6: COMET-22 and COMETKiwi score for four languages sorted by the extend of resource.

process based on STM-generated output using speculative decoding techniques remains future work (Xia et al., 2022; Yang et al., 2023).

5.5 Inherent Language Bias

In this section, we investigate whether the inherent language bias of LLM could be alleviated by combing output from a specialized model. Due to the limited transparency of the GPT-4 model, we turn to a potential indicator, the number of native speakers, to illustrate the extent of a language's resources. We tested on four languages following the setting of SCALE-refine. As shown in Table 6, the performance of few-shot GPT-4 diminishes with the number of native speakers, while our framework, SCALE, consistently and effectively mitigates this language bias, outperforming both few-shot GPT-4 and the supervised NLLB model.

6 Related Work

The use of LLM for translation tasks has garnered significant interest in recent times. Brown et al. (2020) initially demonstrated the efficacy of prompting an LLM with a few examples to achieve noteworthy results, particularly in high-resource languages (Vilar et al., 2023; Lin et al., 2022). Following the release of ChatGPT, several studies have examined its overall translation performance(Jiao et al., 2023; Hendy et al., 2023), along with works focusing on the issue of hallucination (Guerreiro et al., 2023) , literalness (Raunak et al., 2023a), multilinguality (Zhu et al., 2023) and incidental bilingualism problem (Briakou et al., 2023). A comprehensive analysis conducted by Garcia et al. (2023) revealed the unreasonable effectiveness of few-shot LLMs. Furthermore, a diverse range of research has attempted to enhance LLM-based translation systems through cultural awareness (Yao et al., 2023), refinement (Chen et al., 2023), retrieval-augmentation (Cheng et al., 2023b), post-editing (Raunak et al., 2023b), and comparison (Zeng et al., 2023).

Our work also shares similarities with a series of studies aiming to build collaboration between LLMs and other systems. Luo et al. (2023) propose equipping LLMs with a knowledge-guiding module to access relevant information without tuning LLM. Hendy et al. (2023) propose to use Microsoft Translator system as the primary translation system, and then use GPT as a fallback system. Xu et al. (2023) introduce SuperICL and achieve significant improvements in various language understanding tasks.

7 Conclusion

In this paper, we present a novel framework SCALE, which effectively combines the strengths of Large Language Models (LLMs) and compact Specialized Translation Models (STMs) through in-context learning. By providing triplet in-context demonstrations, SCALE unlocks the refinement and pivoting capabilities of LLMs, demonstrated by comprehensive experiments in various settings. Our results offer crucial understanding for subsequent research investigating the possible synergy between LLMs and more specialized models.

8 Limitations

In this paper, we acknowledge the following limitations and strive for improvement as our future work:

(1) While SCALE has demonstrated considerable advancements over both LLMs and STMs across diverse scenarios, our evaluation has predominantly concentrated on the GPT-series as a black-box model. To comprehensively investigate the underlying mechanisms of SCALE, we aim to extend our research to future developments involving powerful multilingual LLMs with fully transparent architectures, weights, and training data distribution.

(2) Although SCALE is the first work to combine LLM and STM into a unified framework, the interaction between these two elements is on the prompting level. Future work will explore more sophisticated integrations, such as applying knowledge distillation from LLMs to STMs, to enhance the synergy between these two components.

(3) The introduction of extended contexts in SCALE is an inevitability that may present significant challenges for systems where response time is critical. Developing strategies to accelerate this process, such as using the output from STMs to perform online speculative decoding, remains an area for further investigation and improvement.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. *URL https://openaipublic. blob. core. windows. net/neuron-explainer/paper/index. html.(Date accessed: 14.05. 2023)*.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in palm's translation capability. *arXiv preprint arXiv:2305.10266*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. *CoRR*, abs/2005.14165.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models.

Xin Cheng, Shen Gao, Lemao Liu, Dongyan Zhao, and Rui Yan. 2022. Neural machine translation with contrastive translation memories. *arXiv preprint arXiv:2212.03140*.

Xin Cheng, Yankai Lin, Xiuying Chen, Dongyan Zhao, and Rui Yan. 2023a. Decouple knowledge from paramters for plug-and-play language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14288–14308, Toronto, Canada. Association for Computational Linguistics.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2023b. Lift yourself up: Retrieval-augmented text generation with self memory.

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

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George F. Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU - neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 46–68. Association for Computational Linguistics.

Xavier Garcia, Yamini Bansal, Colin Cherry, George F. Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. *CoRR*, abs/2302.01398.

Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models.

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. *CoRR*, abs/2302.09210.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-english languages. *arXiv preprint arXiv:1909.09524*.

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 T. 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9019–9052. Association for Computational Linguistics.

Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, and Tong Zhang. 2023. Speciality vs generality: An empirical study on catastrophic forgetting in fine-tuning foundation models.

Ziyang Luo, Can Xu, Pu Zhao, Xiubo Geng, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Augmented large language models with parametric knowledge guiding.

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. *ArXiv*, abs/2303.08774.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, et al. 2023. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*.

Maja Popovic. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 612–618. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan Awadalla. 2023a. Do gpts produce less literal translations?

Vikas Raunak, Matt Post, and Arul Menezes. 2022. SALTED: A framework for SAlient long-tail translation error detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5163–5179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vikas Raunak, Amr Sharaf, Hany Hassan Awadallah, and Arul Menezes. 2023b. Leveraging gpt-4 for automatic translation post-editing.

Ricardo Rei, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luísa 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 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 578–585. 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 2020, Online, November 16-20, 2020*, pages 2685–2702. 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.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Andrea Schioppa, David Vilar, Artem Sokolov, and Katja Filippova. 2021. Controlling machine translation for multiple attributes with additive interventions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6676–6696, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting palm for translation: Assessing strategies and performance.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Heming Xia, Tao Ge, Si-Qing Chen, Furu Wei, and Zhifang Sui. 2022. Speculative decoding: Lossless speedup of autoregressive translation.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. *Advances in neural information processing systems*, 30.

Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. 2023. Small models are valuable plug-ins for large language models. *arXiv preprint arXiv:2305.08848*.

Nan Yang, Tao Ge, Liang Wang, Binxing Jiao, Daxin Jiang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Inference with reference: Lossless acceleration of large language models.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. Empowering llm-based machine translation with cultural awareness. *arXiv preprint arXiv:2305.14328*.

Zheng-Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Indra Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. Bloom+1: Adding language support to bloom for zero-shot prompting.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Tim: Teaching large language models to translate with comparison. *arXiv preprint arXiv:2307.04408*.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. 2023. Mcˆ 2: A multilingual corpus of minority languages in china. *arXiv preprint arXiv:2311.08348*.

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

A Prompt Example

In Table 7, we list the prompt we use for few-shot LLM and in Table 8, for our SCALE framework. We use Chat Markup Language version from Azure to format our prompt[1].

B Data Statistics

We list the detailed data information for SCALE-refine and SCALE-Pivot experiments in Table **??**. The number of dev set is 997 and 1012 for devtest set in flores-200 (NLLB Team et al., 2022).

C Translation Cases

In this section, we list several translation cases from different languages in Figure 6, 7, 8, 9.

D More languages covered with SCALE-pivot

In addition to using Lao as the source language for translations into Assamese, Armenian, Amharic, Xhosa, German, Czech, Bulgarian, and Greek with SCALE-pivot, we demonstrate the versatility of our method by also testing Xhosa as the source language. The results are depicted in Table 10, which exhibit similar patterns with Lao as source languages.

E More languages covered with SCALE-update

In addition to using Xhosa as the source language for translations into English with SCALE-update, we demonstrate the versatility of our method by also testing Lao, Assamese and Amharic as the source language. The results are depicted in Table 11, which exhibit similar patterns with Xhosa as source languages.

[1] `https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/chatgpt?pivots=programming-language-chat-ml`

<table>
<tr><td>Instruction</td><td>< |im_start| >system
Assistant is an intelligent chatbot designed
to help users translate from ${source_language} to ${target_language}
< |im_end| ></td></tr>
<tr><td>Examples</td><td>< |im_start| >user
Source: ${source_1}
Target: ${target_1}
...
Source: ${source_n}
Target: ${target_n}</td></tr>
<tr><td>Input</td><td>Source: ${source}
< |im_end| >
< |im_start| >assistant
Target:</td></tr>
</table>

Table 7: Prompt of Chat Markup Language format for few-shot LLM.

<table>
<tr><td>Instruction</td><td>< |im_start| >system
Assistant is an intelligent chatbot designed
to help users translate from ${source_language} to ${target_language}

Context:
· Assistant would would be given a potentially useful reference answer
from a fine-tuned model
· The number in brackets denotes the confidence score of a fine-tuned model
to generate the token.
< |im_end| ></td></tr>
<tr><td>Examples</td><td>< |im_start| >user
Source: ${source_1}
Potentially useful reference answer 1: ${reference_1}
Potentially useful reference answer 2: ${reference_2}
Target: ${target_1}
...
Source: ${source_n}
Potentially useful reference answer 1: ${reference_1}
Potentially useful reference answer 2: ${reference_2}
Target: ${target_n}</td></tr>
<tr><td>Input</td><td>Source: ${source}
Potentially useful reference answer 1: ${reference_1}
Potentially useful reference answer 2: ${reference_2}
< |im_end| >
< |im_start| >assistant
Target:</td></tr>
</table>

Table 8: Prompt of Chat Markup Language format for SCALE.

code	language	# dev length	# devtest length	script	family	resource
asm_Beng	Assamese	40.55	41.67	Bengali	Indo-European	low
hye_Armn	Armenian	43.91	45.31	Armenian	Indo-European	low
amh_Ethi	Amharic	38.87	39.64	Ge'ez	Afro-Asiatic	low
xho_Latn	Xhosa	35.31	36.37	Latin	Atlantic-Congo	low
uig_Arab	Uyghur	40.77	42.41	Arabic	Turkic	low
khm_Khmr	Khmer	52.77	53.79	Khmer	Austroasiatic	low
npi_Deva	Nepali	34.36	35.48	Devanagari	Indo-European	low
eng_Latn	English	28.99	30.28	Latin	Indo-European	high
deu_Latn	German	37.57	39.16	Latin	Indo-European	high
ces_Latn	Czech	36.63	38.10	Latin	Indo-European	high
bul_Cyrl	Bulgarian	37.99	39.45	Cyrillic	Indo-European	high
rus_Cyrl	Russian	39.42	40.21	Cyrillic	Indo-European	high

Table 9: Data statistics for all the tested languages in the paper.

	Armenian	Assamese	Amharic	Lao	German	Czech	Bulgarian	Greek
# Resource	Low	Low	Low	Low	High	High	High	High
10-shot GPT-4	68.3	65.6	69.3	58.5	74.4	80.8	79.1	78.6
SCALE-pivot	**71.7**	**67.4**	**71.4**	**59.6**	**77.9**	**83.7**	**82.7**	**81.4**

Table 10: Translation results from Xhosa to both low- and high-resource languages, where GPT-4 uses few-shot prompting and SCALE-pivot uses English as the pivot language.

		COMET	BLEURT	spBLEU	chrf++	COMET	BLEURT	spBLEU	chrf++
		xho_Latn				lao_Laoo			
STM	M2M100	68.0	59.0	25.7	46.3	67.8	57.5	13.2	37.9
	NLLB-600M	78.1	70.4	35.0	54.4	84.6	70.3	33.5	55.3
	NLLB-1.3B	80.2	73.1	38.9	57.6	85.8	72.1	36.4	57.7
	NLLB-3.3B	80.7	74.0	40.1	58.3	86.9	73.8	39.4	60.1
10-shot GPT-4		78.8	70.8	34.5	53.3	80.0	63.7	24.5	45.7
SCALE	M2M100	79.1	71.6	34.4	53.7	82.5	67.3	26.8	48.8
	NLLB-600M	81.3	74.6	38.2	57.2	86.3	72.9	34.1	55.4
	NLLB-1.3B	81.9	75.3	39.4	58.3	86.6	73.5	35.5	56.6
	NLLB-3.3B	**82.1**	**75.7**	**40.0**	**58.6**	**87.2**	**74.4**	**38**	**58.5**
		hye_Armn				amh_Ethi			
STM	M2M100	75.9	58.9	23.7	47.9	72.3	54.8	18.5	41.3
	NLLB-600M	86.3	73.4	36.6	58.8	84.7	69.2	30.8	53.6
	NLLB-1.3B	87.7	75.6	40.2	61.4	86.2	71.9	34.0	56.3
	NLLB-3.3B	88.3	77.0	43.0	63.2	86.9	73.6	36.4	58.0
10-shot GPT-4		86.2	73.1	35.6	58.2	83.2	67.3	27.1	48.9
SCALE	M2M100	86.7	74.1	35.8	58.6	84.6	69.7	29.3	51.0
	NLLB-600M	88.1	76.3	39.3	61.0	87.3	74.2	35.3	56.6
	NLLB-1.3B	88.5	77.0	41.2	62.2	87.8	75.1	36.6	57.8
	NLLB-3.3B	**88.8**	**77.8**	**42.3**	**63.1**	**88.0**	**75.7**	**37.6**	**58.5**

Table 11: Results of SCALE-update with different STM (M2M100, NLLB-{600M,1.3B,3.3B}) measured on Xhosa, Lao, Assamese and Amharic to English translation tasks.

SOURCE	बाइसन, एल्क, मूस, भालु र लगभग सबै ठूला जनावरहरूले जस्ता नरम देखिए पनि आक्रमण गर्न सक्छन्।
TARGET	No matter how docile they may look, bison, elk, moose, bears, and nearly all large animals can attack.
MS Translator	Bison, elk, moose, bears, and almost all large animals can attack even if they look soft.
NLLB	The Bible says: ”The one who is walking with wise persons will become wise, but the one who is having dealings with the stupid ones will fare badly”.
GPT-4	Bison, elk, moose, bears, and nearly all large animals, despite appearing gentle, can be aggressive.
SCALE	Bison, elk, moose, bears and nearly all large animals can attack even though they appear docile.

Figure 6: Translation case from Nepali→English.

SOURCE	ভৰি থোৱা ৰিকাবে চলাওঁতাজনৰ ভৰি ৰখাত সহায় কৰে যিটো ঘোঁৰাৰ গাদীৰ দুয়োফালে তললৈ ওলমি থাকে।
TARGET	Stirrups are supports for the rider’s feet that hang down on either side of the saddle.
MS Translator	The legged rickshaw helps to keep the driver’s leg which hangs down on either side of the horse’s mattress.
NLLB	The foot rest helps to keep the rider’s feet which are sloping downwards on both sides of the horse’s saddle.
GPT-4	A heavily loaded Rickshaw helps balance the load by tilting to both sides when going over bumps.
SCALE	The stirrup helps to support the rider’s feet, which are sloping downwards on both sides of the horse’s saddle.

Figure 7: Translation case from Assamese→English.

SOURCE	የዳይኖሶር ላባዎች የዳበረ ራቺስ የሚባል ዘንግ ስለሌለው፣ ነገር ግን ሌሎች የላባ ባህርያት – ባርብስ እና ባርቡልስ – ስላለው ተመራማሪዎች ራቺስ ከእነዚህ ሌሎች ባህርያት የቆየ ዝግመተ ለውጥ ውጤት እንደሆነ ይላሉ።
TARGET	Because the dinosaur feathers do not have a well-developed shaft, called a rachis, but do have other features of feathers — barbs and barbules — the researchers inferred the rachis was likely a later evolutionary development that these other features.
MS Translator	Dinosaur feathers developed because it doesn't have a rod called rachis, but has other feather traits — barbs and barbules — that researchers say is the result of older evolution of rachis from these other traits.
NLLB	dinosaur feathers did not develop a shaft called the rachis, but other feather features, such as barbs and barbels, suggest that the rachis was the result of an earlier evolution of these other features.
GPT-4	As there is no known population of the extinct Laysan Rail on Laysan Island, researchers suggest that the presence of rails on the other islands—Barbados and Barbuda—indicates a prolonged period of isolation and change.
SCALE	Dinosaur feathers did not develop a shaft called the rachis, however, other feather features such as barbs and barbules suggest that the rachis was the result of an earlier evolution of these other features.

Figure 8: Translation case from Amharic→English.

SOURCE	बाइसन, एल्क, मूस, भालु र लगभग सबै ठूला जनावरहरूले जस्ता नरम देखिए पनि आक्रमण गर्न सक्छन्।
TARGET	Auch das Tragen eines Rings ist hilfreich (nur keinen, der zu teuer aussieht
GPT-4	Es gibt eine Chance, dass es genauso verschwindet, wie es aussieht, als ob es einfach verschwindet.
SCALE	Es ist auch nützlich, einen Ring zu tragen, nur scheint der Ring zu teuer zu sein.

Figure 9: Translation case from Lao→German.