Learning-From-Mistakes Prompting for Indigenous Language Translation

You-Cheng Liao, Chen-Jui Yu, Chi-Yi Lin, He-Feng Yun,
Yen-Hsiang Wang, **Hsiao-Min Li**, **Yao-Chung Fan**[*]
Department of Computer Science and Engineering,
National Chung Hsing University, Taiwan
yfan@nchu.edu.tw

Abstract

Using large language models, this paper presents techniques to improve extremely low-resourced indigenous language translations. Our approaches are grounded in the use of (1) the presence of a datastore consisting of a limited number of parallel translation examples, (2) the inherent capabilities of LLMs like GPT-3.5, and (3) a word-level translation dictionary. We harness the potential of LLMs and in-context learning techniques in such a setting for using LLM as universal translators for extremely low-resourced languages. Our methodology hinge on utilizing LLMs as language compilers for selected language pairs, hypothesizing that they could internalize syntactic structures to facilitate accurate translation. We introduce three techniques: *KNN-Prompting with Retrieved Prompting Context, Chain-of-Thought Prompting, and Learning-from-Mistakes Prompting*, with the last method addressing past errors. The evaluation results suggest that, even with limited corpora, LLMs, when paired with proper prompting, can effectively translate extremely low-resource languages.

1 Introduction

In recent years, LLMs have showcased astonishing capabilities in the realm of natural language processing, particularly in tasks like language translation (Zhu et al., 2023), text generation (Yuan et al., 2022), and contextual understanding (Behnia et al., 2022). The robust functionality of these models has led us to reconsider their potential role in indigenous language translation.

In our pursuit to facilitate translations from Chinese to Taiwanese indigenous languages, we leverage the power of LLMs, buttressed by three foundational pillars: the presence of a datastore consisting of a limited number of parallel translation examples, the inherent capabilities of LLMs like GPT-3.5, and the integration of a word-level translation dictionary.

In this paper, we delineate three translation methodologies that build upon each other in a cumulative fashion. Each method represents a layer in our stratified approach, starting from leveraging contextual similarity in KNN-Prompting with Retrieved Prompting Context (RPC) to harnessing the didactic potential of Chain of Thought (CoT) Prompting, and culminating in the Learning-from-Mistakes (LFM) Prompting technique that incorporates feedback mechanisms for continuous improvement. Figure 1 provides an overview of our methodologies, illustrating a step-by-step translation enhancement process designed for the Taiwanese indigenous language context.

This paper is structured as follows: In Section 2, we review the literature and discuss the position of this study. In Section 3, we explore the CoT Prompting methodology, followed by an in-depth analysis of the LFM Prompting approach. In Section 4, we report the evaluation results. Through empirical evaluation and expert reviews, we demonstrate the effectiveness of the proposed methodologies.

2 Related Work

LLMs have exhibited excellent performance in language translation tasks, particularly evident in well-represented source languages like English and Chinese. Despite significant strides in translation performance for these languages, there remains a notable gap in the exploration of LLMs for low-resourced languages or those that have not been pre-trained. This aspect represents an under-explored area within the domain of research.

2.1 Low Resource Translation with LLM

LLMs' effectiveness on various task is primarily attributed to two main properties. Firstly, in-context learning (Brown et al., 2020; Lester et al., 2021) allows the model to learn to solve specific problems by providing a small number of exam-
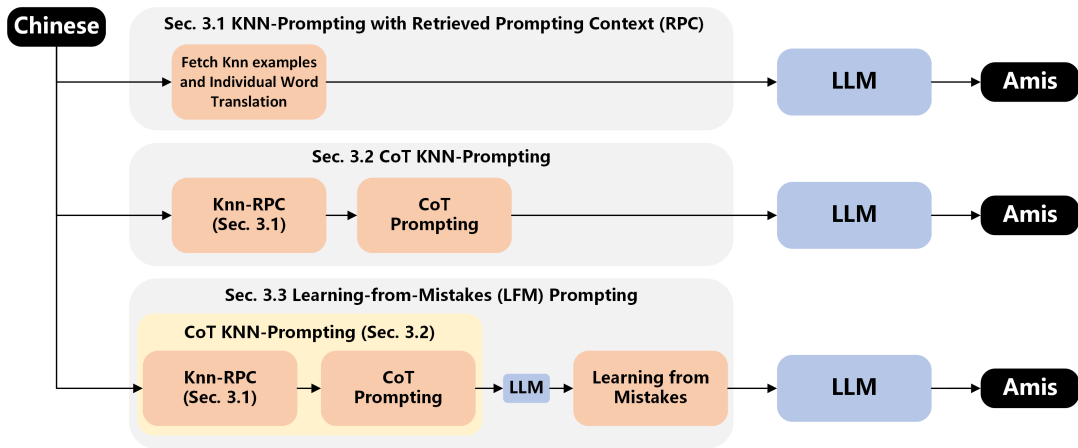
Figure 1: Methodology Overview

ples within the input context. The second one is the ability to follow the instruction (Ouyang et al., 2022; Mishra et al., 2021; Wei et al., 2021), instruction-tuned LLMs can be guiding to solve new task based on text instruction just as the scenario they were trained.

Recently, some research has focused on enhancing these instruction following LLMs through in-context learning, (Nguyen et al., 2023; Ahuja et al., 2023) explores the generation of unsupervised few-shot demonstrations to enhance translation effectiveness in low-resource scenarios. Additionally, (Yao et al., 2023) utilized cultural awareness to optimize alignment in different languages, further augmenting the translation performance of LLMs.

It is noteworthy that, the mentioned works above have focused on low-resource data for LLMs in languages that have been encountered during the pre-training phase. In contrast, our emphasis lies in a scenario where the model has not been previously trained in this specific language. In contrast to conventional approaches, we refrain from training parameters on limited parallel corpora (Gu et al., 2018; Lalrempuii and Soni, 2023). Instead, our goal is to leverage the understanding and reasoning capabilities of LLMs, coupled with the provided data, to accomplish translation tasks for previously unseen languages.

In summary, to the best of our knowledge, no study has delved into the challenges and applications of utilizing LLMs for languages that have not been encountered before.

2.2 Indigenous Language Translation

In the context of preserving and revitalizing indigenous languages, the work by (Zheng et al.,

2022) stands as a notable contribution. Zheng and colleagues introduce the Amis-Mandarin dataset, which includes a parallel corpus comprising 5,751 Amis and Mandarin sentences. This dataset is of particular relevance to our research on translating Chinese sentences into Taiwanese indigenous languages. The Amis-Mandarin dataset provides a valuable resource for studying indigenous language translation. It aligns with the objectives of our study, as it offers a substantial parallel corpus, a fundamental component for training and evaluating translation models. Our research similarly leverages parallel corpora, although we focus on the translation of Chinese into various indigenous languages, including but not limited to Amis. In this study, we conduct experiments on six different indigenous languages.

Furthermore, (Zheng et al., 2022) compile a comprehensive dictionary containing 7,800 unique Amis words and phrases, each accompanied by its Mandarin definition. This lexical resource enhances the utility of their dataset for translation tasks. In our research, we assume the existence of a similar dictionary, emphasizing the importance of word-level translation between Chinese and Taiwanese indigenous languages.

Stap and Araabi (2023) evaluates the translation performance of different systems for Spanish to 11 indigenous languages from South America. The authors find that LLMs like ChatGPT are not yet good at translating into indigenous languages. This is likely due to a number of factors, including the lack of training data for indigenous languages, the complex grammar and sentence structure of indigenous languages, and the difficulty of capturing the nuances of indigenous culture in translation.

2.3 Unveiling LLMs' Proficiency in Tool Usage

In recent research, (Schick et al., 2023) discovered that LLMs exhibit the ability to discern how to employ tools provided by users, including external data. They adeptly combine this external information with their own knowledge to effectively address problem -solving tasks. These investigations delve into the mechanisms of CoT (Inaba et al., 2023) and Self-instruction (Yang et al., 2023) approaches, exploring how these methodologies assist LLMs in comprehending questions and utilizing the tools at their disposal. Additionally, there has been the development of question- answering datasets, such as ToolQA (Zhuang et al., 2023; Inaba et al., 2023), which aimed at faithfully evaluating the ability of LLMs to use external tools for question-answering.

Inspired by these explorations into the understanding and application capabilities of LLMs, we take a similar approach in our method design. We offer KNN examples and word-by-word translation as tools for LLMs to improve their language translation abilities.

2.4 Position of Our Paper

This research stands at the intersection of multiple areas, addressing the challenges of translating into low-resource indigenous languages using LLMs like ChatGPT. While prior works have explored low-resource translation and indigenous language preservation, our study distinguishes itself in two key aspects:

1. **Languages Unseen in Pre-training:** Unlike previous research that has primarily focused on low-resource data for LLMs in languages encountered during pre-training, our work emphasizes the scenario where the model has not been previously trained in the specific target language. We tackle the challenge of translating into languages that lack representation in the model's training data, making our approach more versatile and applicable to a broader range of indigenous languages.

2. **Few-Shot Prompting Techniques:** Our research pioneers the application of few-shot prompting techniques to enhance translation capabilities for indigenous languages. We introduce innovative methods, including *KNN-Prompting with RPC, CoT Prompting, and*

paper	In-context Learning	Fine-tune Parame.	Low-Resource Language	Unseen Language
(Yao et al., 2023)	✓			
(Nguyen et al., 2023)	✓		✓	
(Guerreiro et al., 2023)	✓		✓	
(Gu et al., 2018)		✓	✓	✓
(Lalrempuii and Soni, 2023)		✓	✓	✓
Our work	✓		✓	✓

Table 1: An overview of the existing language translation studies

LFM Prompting, tailored to leverage LLMs' inherent understanding and reasoning abilities. These techniques empower LLMs to effectively tackle low-resource language translation tasks, even when working with limited parallel corpora.

In summary, our paper bridges the gap between LLMs and low-resource indigenous language translation, offering practical and innovative solutions for preserving and revitalizing endangered languages. By exploring the potential of these models in an uncharted linguistic landscape, we provide a fresh perspective and a promising direction for future research in this domain. For clarity, we also compare the related work in Table 1.

3 Methodology

Problem Setting and Assumptions The primary objective of this research is to enable the translation of Chinese sentences into Taiwanese indigenous languages through the utilization of LLMs. In pursuit of this goal, we make the following assumptions for the methods proposed in this study:

- **Datastore of Parallel Corpora:** Our first assumption centers on the availability of a datastore with limited translation examples. Within this datastore, each data entry comprises a pair of sentences: a source sentence in Chinese (the language intended for translation) and a corresponding target sentence in the specific Taiwanese indigenous language. This resource forms the backbone for our translation, facilitating the alignment of linguistic patterns and meanings.

- **Large Language Models:** The cornerstone of our translation methods is the utilization of large pre-trained language models, exemplified by GPT-3.5, as the primary translation engines.

- **Dictionary Existence:** In addition to the aforementioned resources, we introduce an-

other assumption: the existence of a dictionary that spans word-level translations. This dictionary encompasses translations between indigenous language words and their corresponding Chinese counterparts.

Figure 1 outlines our study's methods for enhancing translation in a cumulative manner. The *KNN-Prompting with RPC* method forms the base, merging contextually similar sentences and word translations to inform the LLM's understanding of grammar and context. The *CoT Prompting* adds CoT demonstrations, showing RPC integration for effective translation. The *LFM Prompting* expands upon these with a feedback loop, leveraging previous translation errors to refine outcomes. This progressive strategy not only enhances LLM's translation proficiency but also promotes continual learning and accuracy improvement.

3.1 KNN-Prompting with Retrieved Prompting Context (RPC)

We investigate the application of few-shot learning through the KNN-Prompting concept, as discussed in the works of Shi et al. (Shi et al., 2022) and Xu et al. (Xu et al., 2023). Our approach not only leverages contextually similar examples but also incorporates individual translations for each word in the source language. The methodology unfolds in the following manner:

- When tasked with translating a sentence s, our method initiates by constructing a *Retrieved Prompting Context (RPC)* for s. This context includes:

  - k examples that are contextually analogous to s, selected based on their similarity.
  - Translations for each word in s, sourced from a comprehensive dictionary.

- For instances where direct word equivalents are unavailable, we employ the BERT-base-chinese model as an embedding tool. This model aids in computing similarities to identify the most appropriate substitute words.

- The core principle of our method is to enable the LLM to assimilate the grammatical norms and sentence constructs of the target language. It achieves this through the analysis of the k
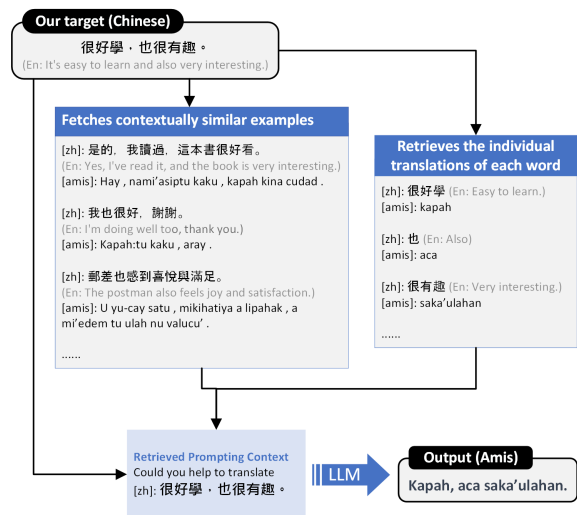

Figure 2: KNN-Prompting with RPC

examples, thereby learning to organize the individually translated words into coherent and grammatically consistent sentences.

For a practical illustration of this process, please see Figure 2, which provides a concrete example of the RPC in action. We also show an example for prompting in Table 5.

3.2 CoT Prompting

In this methodology, we harness the CoT strategy, as delineated by Wei et al. (Wei et al., 2022), to guide the LLM in effectively utilizing the RPC for translating a given sentence s. Specifically, this approach involves the following steps:

- When presented with a sentence s for translation, accompanied by KNN-RPC-prompting inputs (i.e., k contextually similar examples and individual word translations), we further augment the LLM's input with q CoT demonstrations. These demonstrations are designed to illustrate how to use the provided RPC to formulate the final translated sentences.

- An instance showcasing two CoT demonstrations is illustrated in Figure 3. It is important to note that these CoT demonstrations are integrated with the KNN-RPC-prompting inputs to serve as comprehensive prompting material for the LLM.

- For detailed examples of this prompting structure, please refer to Table 6 in the Appendix.
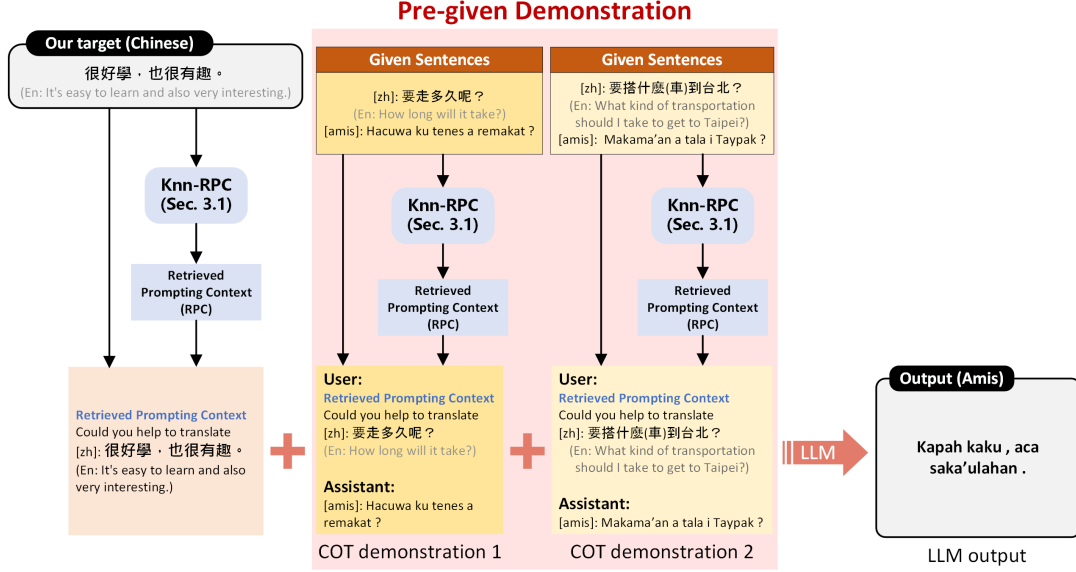
Figure 3: CoT KNN-Prompting: In this example, we have two CoT demonstrations. Note that each CoT demonstration comprises (1) A sample sentence, (2) RPC for the sentence, and (3) The ground-truth sentence. These CoT demonstrations are integrated with the KNN-RPC-prompting inputs to serve as comprehensive prompting material for the LLM.

The overarching aim of this methodology is to empower the LLM with an understanding of the grammatical rules and the proficiency to fully leverage the RPC, including both the retrieved sentences and individual word translations, for producing coherent and accurate translations.

3.3 Learning-from-Mistakes (LFM) Prompting

LFM Prompting is a two-stage approach aimed at enhancing the quality of translations. This method leverages the result from the CoT KNN-Prompting and incorporates a feedback mechanism by conducting trial translation to refine the translation based on past translation errors. The method works in the following phases:

- **Phase 1: Trial Translation with CoT Prompting** When given a sentence s to translate, we start by retrieving q contextually similar sentence pairs from the data store. Each pair (s_{q_i}, t_{q_i}) consists of a Chinese sentence s_{q_i} and its corresponding indigenous sentence t_{q_i}. For each s_{q_i}, we employ CoT KNN-Prompting (the method introduced in Section 3.2) to translate it, resulting in $\hat{t}_{q_i}$. At this stage, we have $(s_{q_i}, t_{q_i}, \hat{t}_{q_i})$. Our approach involves using these results as examples for the LLM to learn from its translation errors and make improvements.

- **Phase 2: Learning from Past Mistakes** The second phase of LFM Prompting introduces a crucial element: the incorporation of past translation errors. Specifically, we treat $(s_{q_i}, t_{q_i}, \hat{t}_{q_i})$ from the Phase 1 as LFM examples. In this phase, we present the LLM with a set of such examples, alongside the translation $\hat{t}$ generated by using CoT KNN-Prompting to translate s. The language model is tasked with refining $\hat{t}$ by considering the error examples in translation. It uses the provided examples of mistranslations to correct and improve the initial translation $\hat{t}$, aligning it more closely with the correct target language structure and meaning.

Furthermore, Figure 4 provides a visual representation of the entire architecture's workflow, illustrating the sequential processes outlined above. We also show a prompting example in Table 7.

4 Experiment

4.1 Model Usage

Utilizing the GPT-3.5-turbo-16k-0613 version with a temperature setting of 0, we employ Sentence BERT (Reimers and Gurevych, 2019) as the embedding model to retrieve k-nearest neighbor sentences. The similarity between sentences is computed using cosine similarity.
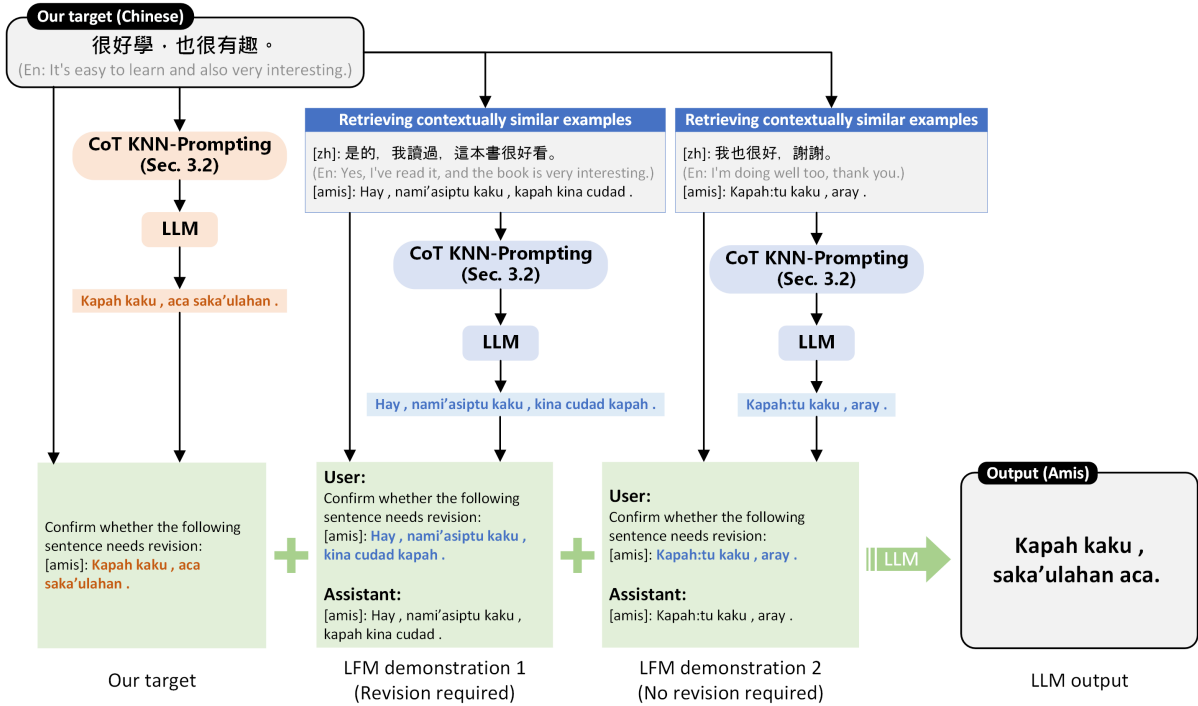
Figure 4: LFM Prompting

	Southern Amis			
Methods	**BLEU1**$_{STD}$	**BLEU2**$_{STD}$	**BLEU3**$_{STD}$	**chrF++**$_{STD}$
Zeroshot	1.0	0.0	0.0	3.9
20-shots	18.0	4.9	1.9	16.3
Knn-Prompting (k=5)	30.1	14.4	6.9	28.1
Knn-Prompting (k=10)	33.3	16.4	8.0	34.2
Knn-Prompting w. RPC (k=5)	$38.2_{2.2}$	$10.5_{1.8}$	$4.3_{1.1}$	$41.2_{1.1}$
Knn-Prompting w. RPC (k=10)	$37.8_{2.2}$	$12.5_{3.2}$	$5.2_{1.9}$	$41.5_{1.6}$
CoT Prompting	$44.4_{1.5}$	$14.3_{0.6}$	$5.9_{1.1}$	$43.5_{0.3}$
LFM Prompting	$44.4_{2.7}$	$17.5_{1.8}$	$8.2_{1.7}$	$44.9_{1.9}$

Table 2: The translation results for **Southern Amis**

4.2 Data Sets

We use the learning materials for various indigenous languages from the 'Klokah' website [1] provided by the Foundation for the Research and Development of Indigenous Languages in Taiwan as our evaluation corpora. Each indigenous group consists of 450 sentences with corresponding Chinese translations and a dictionary of 1000 words (single word translation). For each language, we divide this dataset into two parts:

- **Test Data -** A random selection of 100 sentences was used to evaluate the translation performance of various methods.

- **Reference Data -** The remaining 350 sentences and all dictionaries were used as reference materials for the LLM translation.

---

[1] https://web.klokah.tw/

4.3 Evaluation Results

4.3.1 Automatic Score

We've employed the GPT-3.5-turbo as our foundational language model for translation tests. Initially, we opted for Southern Amis, an indigenous language, as our primary focus, evaluating translation accuracy using the standard BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017) metrics. As depicted in Table 2, the zero-shot translation results indicate the model's limitations in effectively translating this language in the absence of reference data, reflected in BLEU scores nearing zero. However, introducing 20-shot reference data prompts the model to engage in (Agrawal et al., 2022) in-context learning, resulting in a marginal improvement in BLEU scores. This highlights the potential of few-shot learning.

Furthermore, from the results in Table 2, we

Coastal Amis				
Methods	**BLEU1**$_{STD}$	**BLEU2**$_{STD}$	**BLEU3**$_{STD}$	**chrF++**$_{STD}$
Knn-Prompting w. RPC (k=5)	$42.9_{1.8}$	$11.8_{0.9}$	$4.7_{1.1}$	$45.4_{0.8}$
Knn-Prompting w. RPC (k=10)	$43.3_{1.2}$	$13.4_{0.6}$	$5.8_{0.8}$	$44.8_{1.1}$
CoT Prompting	$44.5_{2.8}$	$11.9_{3.0}$	$4.7_{2.3}$	$45.7_{1.6}$
LFM Prompting	$44.1_{2.0}$	$12.6_{2.9}$	$5.7_{2.5}$	$46.1_{1.8}$
Wanda Tayal				
Methods	**BLEU1**$_{STD}$	**BLEU2**$_{STD}$	**BLEU3**$_{STD}$	**chrF++**$_{STD}$
Knn-Prompting w. RPC (k=5)	$41.5_{2.5}$	$13.0_{2.1}$	$4.8_{1.6}$	$42.5_{2.4}$
Knn-Prompting w. RPC (k=10)	$42.1_{2.2}$	$13.6_{2.7}$	$5.7_{1.4}$	$42.8_{2.6}$
CoT Prompting	$46.3_{1.8}$	$14.4_{2.6}$	$5.8_{1.2}$	$44.7_{2.1}$
LFM Prompting	$45.2_{2.0}$	$14.0_{1.7}$	$5.8_{0.9}$	$43.9_{2.0}$
Siji Tayal				
Methods	**BLEU1**$_{STD}$	**BLEU2**$_{STD}$	**BLEU3**$_{STD}$	**chrF++**$_{STD}$
Knn-Prompting w. RPC (k=5)	$44.3_{3.2}$	$14.6_{2.1}$	$4.9_{1.7}$	$39.3_{2.0}$
Knn-Prompting w. RPC (k=10)	$44.4_{3.0}$	$14.5_{2.0}$	$5.4_{2.3}$	$40.9_{1.8}$
CoT Prompting	$47.5_{2.7}$	$16.0_{1.2}$	$5.9_{1.4}$	$41.2_{1.0}$
LFM Prompting	$50.0_{1.2}$	$20.0_{1.4}$	$9.3_{2.0}$	$43.4_{2.0}$
Duda Seediq				
Methods	**BLEU1**$_{STD}$	**BLEU2**$_{STD}$	**BLEU3**$_{STD}$	**chrF++**$_{STD}$
Knn-Prompting w. RPC (k=5)	$45.0_{1.2}$	$16.2_{1.5}$	$5.4_{0.8}$	$38.2_{0.8}$
Knn-Prompting w. RPC (k=10)	$45.7_{1.2}$	$17.1_{1.4}$	$6.7_{1.5}$	$39.3_{1.6}$
CoT Prompting	$46.1_{1.6}$	$17.5_{1.4}$	$6.9_{1.0}$	$38.9_{1.1}$
LFM Prompting	$46.3_{1.5}$	$17.3_{2.1}$	$6.9_{1.4}$	$39.3_{1.2}$

Table 3: The translation results for **Coastal Amis**, **Wanda Tayal**, **Siji Tayal**, and **Duda Seediq**

can observe that using KNN-Prompting by retrieving contextual-relevant examples improves translation quality. We can also observe that utilizing the Chain-of-Thought strategy to guide the LLM also brings an improvement in translation quality, with an increase in BLEU scores from 1 to 3. We also report experiment results with Coastal Amis, Wanda Tayal, Siji Tayal, and Duda Seediq languages. The results are shown in Tables 3.

When comparing models, we use BLEU3 as the main performance metric, as BLEU3 considers 3-gram matches, offering a more holistic view of the quality of translations, particularly in terms of fluency and coherence. In terms of BLEU3 scores, CoT Prompting consistently surpasses the base KNN-Prompting across all languages. This points towards the importance of capturing longer sequences and understanding the grammatical flow of the indigenous languages. We can also see the performance boost when we employ the LFM strategy with CoT in the compared languages.

4.3.2 Qualitive Review by Language Expert

In Table 4 in the appendix, we present the results of an evaluation by a Coastal Amis language expert, assessing translations from Chinese into Coastal Amis using various methods. This offers insights into the effectiveness of these translation strategies and helps us understand their impact on translation quality.

- Initially, the expert demonstrates a preference for translations produced by the LFM method, highlighting its contribution to linguistic precision and affirming the significant role of the LFM phase in enhancing translation quality.

- In the second dialogue, the expert's endorsement of the COT and LFM method suggests that its incorporation can refine the LLM's understanding and conveyance of the target language's nuances.

- There is an identified need for improvement in translating sentence structures, particularly with time adverbs such as ” 非常” (very), ” 很” (very), ” 最” (most), and the placement of temporal terms like ” 今天” (today), ” 明天” (tomorrow) at the sentence's end. The LFM method is anticipated to guide the LLM in learning and internalizing these linguistic patterns, thereby refining the translations. Challenges such as dictionary absences are addressed by seeking synonyms, for instance, substituting ” 專爲” (specially designed for) with ” 最” (most), and ” 南邊” (south) with ” 藍色” (blue). We posit that expanding the dictionary will mitigate such issues, further enhancing translation fidelity.

Overall, the expert's reviews imply that the translation approach integrating the LFM strategy

tends to yield more precise and culturally attuned translations. This suggests that for LLMs translating less-resourced languages, a strategy amalgamating error feedback with accumulative learning might prove more effective. These insights bolster the methodologies delineated in our paper, positing that a stratified and iterative enhancement approach can substantially uplift translation quality, particularly for languages with constrained structural and lexical resources.

5 Conclusion

This study delves into the capabilities of LLMs in translating indigenous languages. Despite a limited datastore of parallel translations, our introduced methodologies: *KNN-Prompting with RPC, CoT Prompting, and LFM Prompting* demonstrate effectiveness in harnessing LLMs for this task. Emphasizing our technical contribution, empirical results highlight the superior performance of the CoT Prompting and LFM strategy over the compared baseline, signifying its adeptness at capturing intricate linguistic nuances and offering an advanced approach to preserving linguistic diversity.

6 Limitations

The strength of this framework lies in its capacity to translate less common, niche languages with a limited number of examples. Nevertheless, several challenges were encountered during the experiments. For example, in the case of the Southern Amis language, the term 'we' can be translated as 'kami' or 'niyam,' among other options. Determining whether these terms carry subtle distinctions in meaning or are interchangeable necessitates the expertise of native speakers. Moreover, the use of the BLEU metric provides only one standardized answer, which may not consistently align with the actual context.

Furthermore, within the LFM context, structural or grammatical corrections are solely guided by prior examples, as the language model itself lacks the capability for independent reasoning and adjustment. Therefore, achieving significant breakthroughs in effectiveness remains a challenge. Finally, while our methods have demonstrated data-driven enhancements, they do not fully address the issue of insufficient few-shot data resulting in inconsistent translation outcomes. Further research and innovation are essential in addressing this matter.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Rouzbeh Behnia, Mohammadreza Reza Ebrahimi, Jason Pacheco, and Balaji Padmanabhan. 2022. Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 560–566. IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal neural machine translation for extremely low resource languages. *arXiv preprint arXiv:1802.05368*.

Nuno M Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023. Hallucinations in large multilingual translation models. *arXiv preprint arXiv:2303.16104*.

Tatsuro Inaba, Hirokazu Kiyomaru, Fei Cheng, and Sadao Kurohashi. 2023. Multitool-cot: Gpt-3 can use multiple external tools with chain of thought prompting. *arXiv preprint arXiv:2305.16896*.

Candy Lalrempuii and Badal Soni. 2023. Investigating unsupervised neural machine translation for low-resource language pair english-mizo via lexically enhanced pre-trained language models. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(8):1–18.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2023. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. *arXiv preprint arXiv:2306.11372*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Weijia Shi, Julian Michael, Suchin Gururangan, and Luke Zettlemoyer. 2022. Nearest neighbor zero-shot inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3254–3265.

David Stap and Ali Araabi. 2023. Chatgpt is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Benfeng Xu, Quan Wang, Zhendong Mao, Yajuan Lyu, Qiaoqiao She, and Yongdong Zhang. 2023. k nn prompting: Beyond-context learning with

calibration-free nearest neighbor inference. *arXiv preprint arXiv:2303.13824*.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023. Gpt4tools: Teaching large language model to use tools via self-instruction. *arXiv preprint arXiv:2305.18752*.

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

Xingdi Yuan, Tong Wang, Yen-Hsiang Wang, Emery Fine, Rania Abdelghani, Pauline Lucas, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2022. Selecting better samples from pre-trained llms: A case study on question generation. *arXiv preprint arXiv:2209.11000*.

Francis Zheng, Edison Marrese-Taylor, and Yutaka Matsuko. 2022. A parallel corpus and dictionary for amis-mandarin translation. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 79–84.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *arXiv preprint arXiv:2306.13304*.

Appendix

Coastal Amis language expert reviews translation results
[zh]: 你喜歡看書嗎？(En: Do you like reading books?) [CoT]: Kiso (你, you) maolah (喜歡, like) minengneng (看望, visit) a cudad (書, book) haw ? [LFM]: Kiso (你, you) maolah (喜歡, like) cudad (書, book) minengneng (看望, visit) haw ? [language expert]: Maolah (喜歡, like) kiso (你, you) minengneng (看望, visit) to codad (書, book) haw?
[zh]: 是的，我很喜歡。(En: Yes, I really like it.) [CoT]: Hai (是的, yes) , maolah (很喜歡, really like) to kako (我, I) . [LFM]: Hai (是的, yes) , maolah (很喜歡, really like) to kako (我, I) . [language expert]: Hai (是的, yes), ma'olah (很喜歡, really like) kako (我, I).
[zh]: 你會織布嗎？(En: Do you know how to weave?) [CoT]: Mafana' (會, know how to) kiso (你, you) a miteno'oy (織布, Weaving fabric) ? [LFM]: Mafana' (會, know how to) kiso (你, you) miteno'oy (織布, Weaving fabric) ? [language expert]: Mafana' (會, know how to) kiso (你, you) miteno'oy (織布, Weaving fabric) haw?
[zh]: 我最喜歡夏天。(En: I like summer the most.) [CoT]: Maolah (很喜歡, really like) kako (我, I) patodongan (專爲, Specially designed for) ko kaciherangan (夏天, summer) . [LFM]:Maolah (很喜歡, really like) kako (我, I) kaciherangan (夏天, summer) patodongan (專爲, Specially designed for) . [language expert]: O kaciherangan (夏天, summer) ko kaolahan (最喜愛的, favorite) ako(我, I).
[zh]: 我最喜歡藍色。(En: I like blue the most.) [CoT]: Maolah (喜歡, like) kako (我, I) patodongan (專爲, Specially designed for) satimolan (南邊, south) . [LFM]: Maolah (喜歡, like) kako (我, I) satimolan (南邊, south) patodongan (專爲, Specially designed for) . [language expert]: O langdaway (藍色, blue) a cengel (顏色, color) ko kaolahan (最喜歡, favorite) ako (我, I).
[zh]: 是的，天氣非常好。(En: Yes, the weather is very nice.) [CoT]: Hai (是的, yes) , romi'ad (天氣, weather) tada (非常, very) nga'ay (好, good) . [LFM]: Hai (是的, yes) , romi'ad (天氣, weather) tada (非常, very) nga'ay (好, good) . [language expert]: Hai (是的, yes), fangcal (很好, very good) ko romi'ad (天氣, weather) anini (今天, today).

Table 4: Result of Language Expert Review

Knn-Prompting with RPC

You are an Amis language translator. The followings are some [zh] to [amis] examples.
Chinese: 是的，我讀過，這本書很好看。(English: Yes, I've read it, and the book is very interesting.)
[Amis]: Hay, nami'asiptu kaku, kapah kina cudad.
[zh]: 我也很好，謝謝。(English: I'm doing well too, thank you.)
[Amis]: Kapah:tu kaku, aray.
[zh]: 郵差也感到喜悅與滿足。(English: The postman also feels joy and satisfaction.)
[Amis]: U yu-cay satu, mikihatiya a lipahak, a mi'edem tu ulah nu valucu'.
  :
  :
[zh]: 很好學 (English: Easy to learn.)
[Amis]: kapah
[zh]: 也 (English: Also)
[Amis]: aca
[zh]: 很有趣 (English: Very interesting).
[Amis]: saka'ulahan
Based on the above examples. Could you help to translate [zh]: 很好學，也很有趣. (English: It's easy to learn and interesting)

Table 5: Simplified Example for Knn-Prompting with RPC

CoT Demonstration 1

You are an Amis language translator. The followings some [zh] to [amis] examples.
[zh]: 現在幾點鐘？(En: What time is it now?)
[Amis]: Pina’ay ku tuki anini ?
[zh]: 喔！她什麼時候回來呢？(En: Oh! When is she coming back?)
[Amis]: A, a hacuwa cira a taluma’ ?
  :
  :
[zh]: 要 (En: Need)
[Amis]: aw
[zh]: 走 (En: Go)
[Amis]: rakat
[zh]: 幾天 (En: How many days.)
[Amis]: kapina a remi’ad
[zh]: 呢 (En: Question particle)
[Amis]: saw
Based on the above examples. Could you help to translate [zh]: 要走多久呢？
[Assistant:] Hacuwa ku tenes a remakat ?

CoT Demonstration 2

You are an Amis language translator. The followings some [zh] to [amis] examples.
[zh]: 今年，我們伯伯全家人從台北搭車回來。(En: This year, our uncle's entire family came back from Taipei by car.)
[Amis]: Anini a miheca, makakarireng a taluma’ ku vaki niyam atu wawa nira namaka Taypey.
[zh]: 火車比較快。(En: The train is faster.)
[Amis]: U silamalay ku kalamkamay.
  :
  :
[zh]: 要 (En: Need)
[Amis]: aw
[zh]: 搭什麼 (En: Take what.)
[Amis]: Makama’an
[zh]: 公車 (En: Bus)
[Amis]: vasu
[zh]: 到 (En: Arrive)
[Amis]: tangasa
[zh]: 台北 (En: Taipei)
[Amis]: Taypak
Based on the above examples. Could you help to translate [zh]: 要搭什麼（車）到台北？
[Assistant:] Makama’an a tala i Taypak ?

CoT Prompting

You are an Amis language translator. The followings are some [zh] to [amis] examples.
Chinese: 是的，我讀過，這本書很好看。(English: Yes, I've read it, and the book is very interesting.)
[Amis]: Hay, nami’asiptu kaku, kapah kina cudad.
[zh]: 我也很好，謝謝。(English: I'm doing well too, thank you.)
[Amis]: Kapah:tu kaku, aray.
[zh]: 郵差也感到喜悅與滿足。(English: The postman also feels joy and satisfaction.)
[Amis]: U yu-cay satu, mikihatiya a lipahak, a mi’edem tu ulah nu valucu’.
[zh]: 很好學 (English: Easy to learn.)
[Amis]: kapah
[zh]: 也 (English: Also)
[Amis]: aca
[zh]: 很有趣 (English: Very interesting).
[Amis]: saka’ulahan
Based on the above examples. Could you help to translate [zh]: 很好學，也很有趣. (English: It's easy to learn and interesting)

Table 6: Example for Simplified CoT KNN-Prompting

LFM Prompting Example

Please analyze the differences between [Your Answer] and [Correct Answer] results.
[zh]: 是的，我讀過，這本書很好看。(En: Yes, I've read it, and the book is very interesting.)
[Your Answer]:Hay , nami' asiptu kaku , kina cudad kapah .
[Correct Answer]: Hay , nami' asiptu kaku , kapah kina cudad .
Please analyze the differences between [Your Answer] and [Correct Answer] results.
[zh]: 我也很好，謝謝。(En: I'm doing well too, thank you.)
[Your Answer]:Kapah:tu kaku , aray .
[Correct Answer]: Kapah:tu kaku , aray .

You are an Amis language translator. The followings some [zh] to [amis] examples.
[zh]: 是的，我讀過，這本書很好看。(En: Yes, I've read it, and the book is very interesting.)
[amis]:Hay , nami' asiptu kaku , kapah kina cudad .
[zh]: 我也很好，謝謝。(En: I'm doing well too, thank you.)
[amis]:Kapah:tu kaku , aray .
[zh]: 郵差也感到喜悦與满足。(En: The postman also feels joy and satisfaction.)
[amis]:U yu-cay satu , mikihatiya a lipahak , a mi' edem tu ulah nu valucu' .
...
[zh]: 很好學 (En: Easy to learn.)
[amis]: kapah
[zh]: 也 (En: Also)
[amis]: aca
[zh]: 很有趣 (En: Very interesting.)
[amis]: saka' ulahan
Check whether the following sentence needs revision:
[zh]: 很好學，也很有趣。(English: It's easy to learn and interesting)
[Your Answer]:Kapah kaku , aca saka' ulahan .
[Correct Answer]:

Table 7: Example for Simplified LFM Prompting. Note that as introduced in the LFM method, when given a sentence s (i.e., 很好學，也很有趣) to translate, we start by retrieving q contextually similar sentence pairs from the data store and use CoT KNN prompting to obtain trial translation results (the sentence followed [Your Answer]) and also the correct answer for enabling LFM.