Bridging the Gap: Adapting LLMs for Southeast Asian Low-Resource Machine Translation via Hierarchical Dynamic Retrieval and Matching

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) has proven its effectiveness in enhancing the generation capabilities of large language models (LLMs) for various natural language processing tasks. However, its ability in low-resource machine translation drops sharply due to the noise interference caused by the semantic mismatch between retrieved content and translation requirements. To alleviate this drawback, we propose a novel hierarchical dynamic retrieval and matching approach for Southeast Asian low-resource machine translation. First, we construct a hierarchical index structure that utilizes high-frequency word statistics as key indices based on an existing parallel corpus, associating bilingual short and long sentence pairs. Second, we dynamically match words between the source sentence and the hierarchical index structure to retrieve all associated short and long bilingual sentence pairs. Meanwhile, we rerank the candidate samples by computing cross-lingual semantic similarity between the source sentence and the retrieved pairs. Finally, the sample with the highest semantic similarity is integrated into the prompt to guide LLMs in generating more accurate translations. Experimental results show that our approach outperforms mainstream machine translation systems without fine-tuning LLM parameters. Detailed analysis indicates that our method precisely matches fine-grained semantic information, thus reducing noise interference and improving low-resource translation performance.

1 Introduction

Retrieval-Augmented Generation (RAG) enhances large language models (LLMs) by dynamically retrieving contextually relevant information from external knowledge bases to refine generation fidelity(Chen et al., 2024b; Asai et al., 2023). While conventional LLMs exhibit formidable text generation capabilities through pretraining on massive corpora with billions of parameters, their performance is constrained by static knowledge boundaries and temporal data limitations, often leading to factual inaccuracies or semantic inconsistencies in domain-specific applications(Lewis et al., 2020; Huang and Huang, 2024). The RAG framework addresses these constraints through an adaptive knowledge retrieval mechanism that supplements real-time contextual knowledge without requiring parameter updates, thereby improving both output accuracy and domain adaptability.

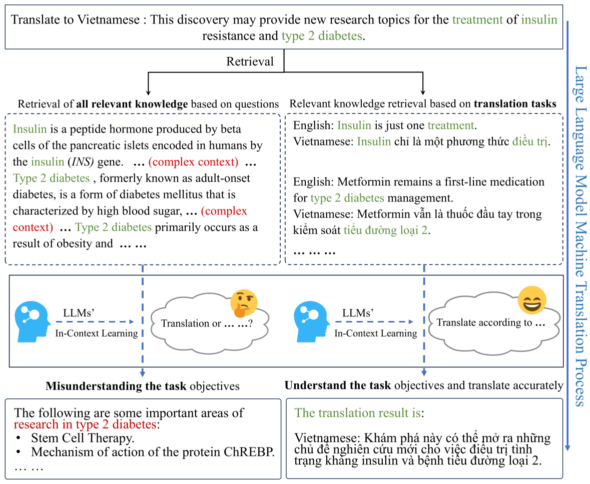

Figure 1: Possible scenarios of LLMs using RAG for contextual hinting in low-resource machine translation tasks.

Although RAG has proven effective in enhancing text generation through contextual learning with LLMs, several critical challenges persist in practical implementations. First, as demonstrated by Zhu et al. (2024a) and Xu et al. (2024), excessive contextual prompts introduce noise and misinformation that significantly impair LLMs' comprehension and generation capabilities. Second, in machine translation tasks, weak semantic relevance between retrieved content and source text often leads to imprecise outputs that mismatch target language contexts(Min et al., 2022). This issue is exacerbated in low-resource language scenarios where conventional retrieval methods struggle to identify semantically aligned sentences from sparse parallel corpora(Fig.1). Furthermore, the scarcity of training data for low-resource language pairs hinders multilingual models' ability to achieve precise text alignment and capture cultural nuances in target languages(Hendy et al., 2023; Alam et al., 2024). While RAG substantially improves LLM performance across general tasks, optimizing knowledge base construction and retrieval strategies remains pivotal for advancing its effectiveness in low-resource machine translation.

To address these challenges, we propose a RAG-enhanced translation framework with precision retrieval from external parallel corpora, specifically targeting low-resource Chinese-to-Southeast Asian language pairs (Vietnamese, Burmese, Indonesian, Malay). Our methodology first constructs a cleaned bilingual corpus through word frequency analysis and filtering of stopwords/numerical tokens. The corpus is subsequently dynamically segmented into short and long sentence pairs based on sentence length, then organized hierarchically into a tripartite retrieval tree using Chinese character frequency statistics. To ensure semantic coherence, we employ a cross-lingual sentence encoder for vector representation of sentence pairs and compute cross-lingual cosine similarity scores between bilingual embeddings. During inference, the retrieval tree efficiently retrieves contextually optimal prompt pairs (short and long) as dynamic context for large language models, enhancing translation accuracy and target-language appropriateness.

This method effectively organizes and optimizes the storage and retrieval of corpora by constructing a tree-based retrieval structure with words as the root nodes. This structure ensures high relevance of the retrieved prompts while capturing the critical keywords of the sentences to be translated, significantly reducing the impact of irrelevant information on the translation process. The main contributions of this work are summarized as follows:

1. Integrates RAG into low-resource LLM-based machine translation, significantly enhancing prompt relevance and domain-specific adaptation through dynamic knowledge injection.

2. Achieves translation quality parity with mainstream NMT systems via retrieval-augmented context prompting, while preserving LLM parameters without model fine-tuning.

3. Systematically evaluates context window scaling effects (128-640 tokens) on low-resource MT performance, revealing distinct optimal length ranges for different Southeast Asian languages.

2 Related Work

Neural Machine Translation. Neural Machine Translation represents a fundamental technology in natural language processing, evolving from rule-based systems to the current data-driven paradigm(Och and Ney, 2002; Koehn et al., 2003). Transformer-based architectures establish themselves as the dominant approach(Vaswani et al., 2017), leveraging self-attention mechanisms to achieve end-to-end semantic modeling and significantly improving translation fluency and cross-lingual consistency for high-resource language pairs(Bahdanau et al., 2014; Devlin et al., 2014). However, these systems face critical limitations in low-resource scenarios due to their dependence on large-scale parallel corpora. Data sparsity compromises the models' ability to generalize target language patterns, while the long-tail distribution of linguistic representations further degrades translation quality(Ranathunga et al., 2023).

Recent advances address these challenges through two primary strategies. Data augmentation techniques generate pseudo-parallel corpora to mitigate training data scarcity(Lample et al., 2017; Prabhumoye et al., 2018; Imankulova et al., 2019; Ouyang et al., 2020). Meanwhile, large multilingual language models create shared cross-lingual semantic spaces, enhancing transfer learning capabilities(Radford et al., 2018; Touvron et al., 2023). Notable implementations include the NLLB model, which supports direct translation across 200+ languages(Costa-Jussà et al., 2022), and M2M-100(Fan et al., 2021), eliminating English-centric pivoting for non-English language pairs. Parameter-efficient fine-tuning methods like LoRA(Hu et al., 2022) and Adapter(Houlsby et al., 2019) further reduce dependency on annotated data while maintaining model performance.

Nevertheless, significant challenges persist for Southeast Asian low-resource languages. The English-dominated nature of pretraining data introduces inherent biases, and the extreme scarcity of parallel corpora compounds alignment difficulties(Bender et al., 2021; Winata et al., 2021). While current multilingual systems provide broad language coverage, their translation quality for low-resource pairs remains substantially inferior to high-resource scenarios, highlighting the need for continued methodological innovation(Le Scao et al., 2023).

Retrieval-Augmented Generation. RAG provides an innovative path to address the static knowledge limitations of traditional LLMs by dynamically integrating external knowledge retrieval and the reasoning capabilities of generation models. The Navie RAG framework follows the standard "retrieval-generation" unidirectional pipeline design(Ma et al., 2023), but its performance is limited by the semantic misalignment between retrieval noise and generation targets, which easily causes the output to deviate from the real context. To this end, Advanced RAG introduces a multi-stage optimization mechanism, such as gradually refining the query intent through multiple rounds of iterative retrieval(Wang et al., 2024; Sawarkar et al., 2024), or re-ranking the relevance of retrieval results to screen high-confidence content(Feng et al., 2024; Yoon et al., 2024), which significantly improves the efficiency of knowledge fusion. Furthermore, Modular RAG supports flexible configuration of heterogeneous components by decoupling the architecture design of the retriever and the generator(Gao et al., 2024; Wang et al., 2023). The retrieval side can integrate dense vector retrieval and sparse keyword matching strategies, while the generation side can adapt to pre-trained models of different sizes and achieve task customization by combining domain fine-tuning.

While RAG has demonstrated considerable success in general text generation, its application to low-resource machine translation presents unique challenges. The scarcity of parallel corpora hinders cross-lingual alignment, while morphological and syntactic complexities in target languages exacerbate semantic discrepancies. Although research indicates that language-aware contextual prompts can enhance translation quality(Puduppully et al., 2023; Zhu et al., 2024b), standard RAG frameworks face inherent limitations in multilingual settings. Key issues include inadequate coverage of low-resource languages in existing retrieval models and the introduction of noise when directly incorporating retrieved texts(Jiang et al., 2023; Shi et al., 2023), compounded by the absence of language-specific filtering mechanisms. These challenges underscore the need for specialized retrieval architectures tailored to low-resource translation scenarios.

3 Methodology

This study proposes a retrieval-enhanced method for low-resource machine translation using large language models. The primary objective is to leverage high-quality prompt information from an external parallel corpus to improve the translation model's performance in low-resource language tasks. Specifically, given a prompt retrieval dataset $\mathbb{D} = \{(x_i, y_i)\}_{i=1}^{|\mathbb{D}|}$, where (x_i, y_i) represents a pair of parallel sentences from the source language x_i to the target language y_i, the dataset is stored in a tree-structured database for efficient retrieval. For a given sentence S_t to be translated, a subset of relevant parallel sentence pairs $\mathbb{D}' = \{D_m\}_{m=1}^{|M|}$ is retrieved from $\mathbb{D}$ where each D_m $(1 \le m \le M)$ serves as a translation prompt closely related to S_t. Based on these retrieved sentence pairs and the input sentence, a translation prompt template is constructed as:

$$P_t = \{(D_m, S_t)\,|m = 1, 2, ..., |M|\}$$

these templates are then used to generate contextually relevant prompts, effectively enhancing the translation model's performance. Figure 2 provides an overview of the model framework and the translation process.

3.1 Word Statistics

Before building the Parallel Sentence Pair Retrieval Tree(PSP-RT) structure, we need to clean and count the bilingual parallel corpus in the Chinese part to ensure the corpus quality and retrieval efficiency. First, we collected a large-scale bilingual parallel corpus from Chinese to low-resource languages and preprocessed the corpus, including converting traditional Chinese into simplified Chinese, removing redundant characters, bilingual sentence pairs with abnormal lengths, and irrelevant punctuation.

During the word-counting process, this study adopts the concept of the Inverted File Index (IVF) to efficiently establish the mapping relationship between words and the index set of the sentences in which they appear(Babenko and Lempitsky, 2014). Let $D_x = \{x_i|\,(x_i, y_i \in D)\}$ denote the set of all Chinese sentences. For any word $w \in D_x$, its sentence index set $I(w)$ is defined as:

$$I\,(w) = \{i|w \in x_i, x_i \in D_x\}$$

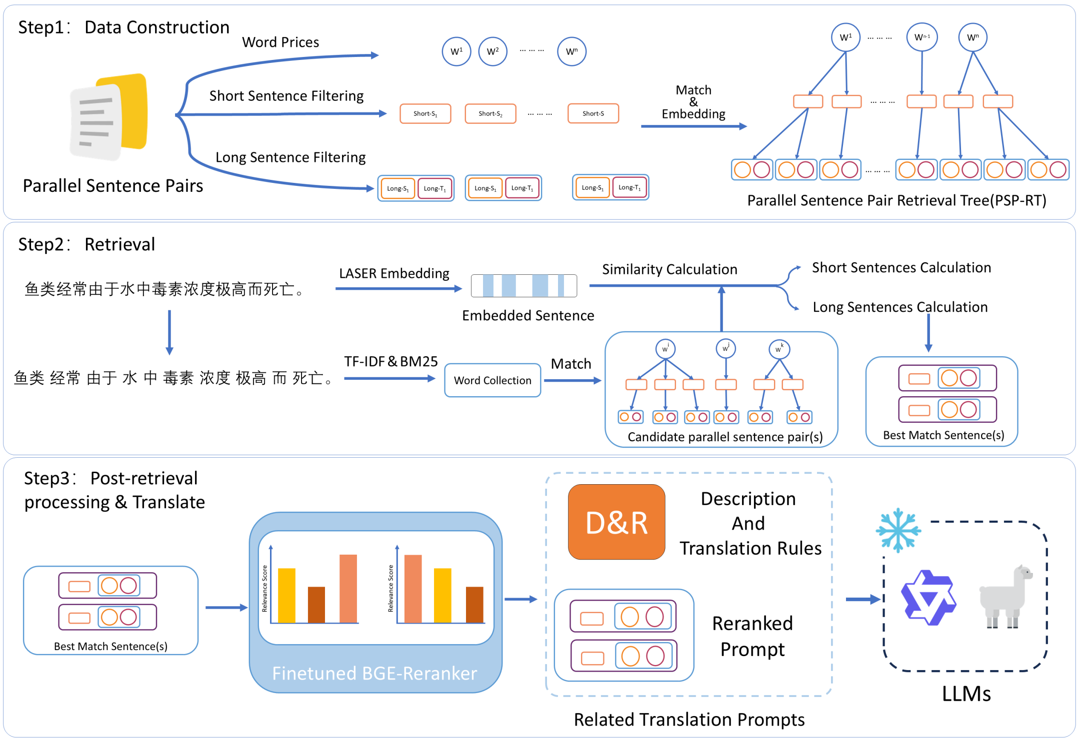

Figure 2: Framework of our proposed approach.

where i represents the index of the i-th sentence containing the word w.

To select the most representative words as root nodes for the PSP-RT in the subsequent construction process, Chinese stop words are removed, and all remaining words are sorted in ascending order based on the size of their sentence index sets. The sorting rule is as follows:

$$w_1, w2, ..., w_n. \quad subject\ to$$
$$|I(w_1)| \leq |I(w_2)| \leq ... \leq |I(w_n)|$$

this sorting rule prioritizes low-frequency words as root nodes, as they typically exhibit greater discriminative power and are associated with fewer sentences. This approach effectively narrows the scope of candidate sentence pairs during retrieval. In contrast, high-frequency words, which appear in a large number of sentences, can lead to an over-concentration of Chinese-to-low-resource-language sentence pairs if sorted by descending order of sentence coverage. Such concentration reduces the diversity of retrieval results and diminishes the relevance of prompt information.

3.2 Parallel Sentence Pair Insertion

Long and Short Sentence Pairs Division. Prior to constructing the PSP-RT, the bilingual parallel corpus is divided into short sentence pairs and long sentence pairs with a ratio of 1:3. Short sentence pairs, owing to their brevity, are more efficient for rapid matching with individual words. In contrast, long sentence pairs, when matched with short sentence pairs, leverage both the distributional information of words within sentences and the semantic similarity at the sentence level. This division strategy is designed to enhance both the efficiency of PSP-RT construction and the quality of retrieval outcomes. This dual consideration ensures higher precision in retrieval results while maintaining semantic diversity, thereby improving the robustness and reliability of the retrieval process.

Short Sentence Pair Insertion. The insertion process for short sentence pairs begins by filtering the words in the sentence using the TF-IDF algorithm, which identifies candidate words with high relevance to the given sentence. Subsequently, the BM25 algorithm(Robertson et al., 2009) is applied to compute the matching scores between the candidate words and the sentence, selecting the word with the highest score as the root node of the PSP-RT. Based on these computations, the short sentence pair is inserted into the appropriate node of the PSP-RT, while simultaneously recording the word indices and sentence pair information.

To further enhance semantic retrieval capabilities, the LASER model(Artetxe and Schwenk,

2018) is employed to generate bilingual embeddings for the short sentence pairs. These semantic representations are stored under the corresponding PSP-RT nodes, providing a robust foundation for subsequent semantic-level matching and retrieval tasks.

Long Sentence Pair Insertion. The insertion process for long sentence pairs builds on the semantic matching capabilities established with short sentence pairs, further enhancing the PSP-RT's structure. For each long sentence pair, the TF-IDF and BM25 algorithms are first employed to calculate its relevance to each word in the PSP-RT. Words with the highest relevance are selected as candidate root nodes. Based on these nodes, all short sentence pairs stored beneath them are retrieved as potential matching targets.

Next, the LASER model is utilized to perform bilingual encoding of the long sentence pair, resulting in semantic vectors for the source and target languages, denoted as x_L and y_L respectively. Similarly, the semantic vectors of the retrieved short sentence pairs are extracted and recorded as x_S and y_S. The semantic similarity between the long and short sentence pairs is then computed using cosine similarity, forming a 2×2 similarity matrix S,

$$S = \begin{bmatrix} sim(x_L, x_S) & sim(x_L, y_S) \\ sim(y_L, x_S) & sim(y_L, x_L) \end{bmatrix}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity function. To comprehensively evaluate the semantic similarity across all matching paths, the Frobenius norm of the similarity matrix is calculated, serving as the basis for determining the insertion position of the long sentence pair. The comprehensive score is defined as:

$$Score\,(S) = ||S||_F = \sqrt{\sum_{i=1}^{2}\sum_{j=1}^{2} S_{ij}^2}$$

where $||S||_F$ represents the Frobenius norm of matrix S. By integrating both word-level and sentence-level semantic information, this method mitigates the issue of aggregation caused by over-reliance on high-frequency words, enhancing the diversity and relevance of the bilingual sentence pair retrieval process.

3.3 Post-Retrieval Processing for Translation

This section outlines the retrieval and post-processing workflow for candidate sentence pairs related to the sentence to be translated. Initially, relevant long and short sentence pairs are extracted from the constructed PSP-RT through the retrieval process. Subsequently, the input sentence is encoded and cosine similarity scores are computed between the input and both the source and target language parts of the retrieved sentences. The sentence pairs are then preliminarily ranked based on the average of these scores. Since the initial retrieval may yield a large number of candidate sentence pairs, some of which may have low relevance or contain redundant information, directly using these pairs as prompts could compromise the accuracy of the translation results. To address this, we utilize a fine-tuned reranker model to further refine and reorder the candidate sentence pairs, ensuring that only highly relevant prompts are selected.

The reranker model prioritizes the candidate pairs based on their relevance scores, selecting the most relevant bilingual sentence pairs as prompts for input into the large language model. This strategy ensures that the prompts are semantically aligned with the sentence to be translated, improving translation accuracy and contextual consistency. During the translation generation process, the parameters of the large language model remain frozen and are solely used for generating translations without involving any parameter fine-tuning. By combining retrieval with re-ranking, this approach provides the large language model with high-quality contextual prompts, effectively enhancing translation accuracy while maintaining computational efficiency.

4 Experiments

4.1 Experimental Setup

Datasets. To construct a parallel corpus for building the PSP-RT from Chinese to low-resource languages, we aggregate multilingual parallel data from three major public sources: CCMatrix(Schwenk et al., 2019), NLLB, and the Asian Language Treebank (ALT)(Thu et al., 2016). During preprocessing, we first filter out sentence pairs with anomalous lengths or redundant characters, while converting all traditional Chinese characters to simplified form. We then employ both langdetect[1] and polyglot[2] for rigorous language identification, eliminating any pair of sentences with inconsistent source-target language labeling to ensure

[1] https://github.com/Mimino666/langdetect
[2] https://github.com/aboSamoor/polyglot

corpus quality. Table 1 shows the experimental usage data. For translation evaluation, we adopt the Flores-200 benchmark dataset, a widely recognized multilingual evaluation resource for machine translation systems.(Costa-Jussà et al., 2022).

Table 1: The number of bilingual parallel corpora used to construct the Parallel Sentence Pair Retrieval Tree.

Language Pair	Raw Data	Usage Data
Chinese-Vietnamese	2,701,926	959,995
Chinese-Burmese	630,000	442,622
Chinese-Indonesian	3,930,249	1,062,167
Chinese-Malaysian	2,082,910	646,646

Baselines and Evaluation Metrics. In the experimental phase, we select Qwen 2.5-7B(Yang et al., 2024) and Llama 3.1-8B(Grattafiori et al., 2024) as baseline machine translation models, with NLLB-200-distilled-600M serving as a strong multilingual baseline. We systematically evaluate the translation performance of these models under default settings, along with their variants augmented by the Naive RAG retrieval method, and compare them against our proposed approach. For evaluation, we used a composite metric framework comprising spBLEU (SentencePiece-based BLEU)(Goyal et al., 2022) and COMET(Rei et al., 2022). The spBLEU metric utilizes the SentencePiece tokenizer to enforce a unified subword segmentation scheme, thereby eliminating tokenization discrepancies that skew traditional BLEU scores in cross-lingual evaluations. The COMET metric employs the Unbabel/wmt22-comet-da model, which utilizes multilingual BERT to assess translation quality across three dimensions: semantic coherence, lexical appropriateness, and contextual consistency.

Model fine-tuning. In order to enhance the reranking model's adaptability to large language models and their translation tasks, we use the fine-tuned BGE-reranker-v2-m3 model (Li et al., 2023) to reorder the retrieved bilingual prompt sentence pairs. To facilitate effective training, we construct a fine-tuning dataset comprising 200K samples. Each sample in this dataset consists of a query, which is a Chinese source sentence, a positive example corresponding to its target language translation, and fifteen negative examples generated by randomly sampling other target language sentences from the dataset. During training, we adopt a dynamic learning rate strategy, initializing the learning rate at 2e-5 and reducing it by a factor of 0.7 after each training epoch.

Implementation Details. To ensure the provision of sufficient hints during the retrieval process and prevent the performance of the large language model from being constrained by an insufficient number of relevant sentence pairs, we design a default retrieval path based on the ALT corpus. In practice, when the number of bilingual sentence pairs retrieved is fewer than 5, we supplement the results with semantically similar content from the default retrieval path. This approach ensures the adequacy of the hints and enhances the overall quality of the translation task.

4.2 Experimental Results

To systematically validate the effectiveness of the proposed method, we design a comparative experimental protocol: first evaluating the baseline performance of Qwen 2.5-7B and Llama 3.1-8B under zero-shot settings; then selecting five high-quality bilingual prompt texts from the ALT dataset following (Hendy et al., 2023); while adopting the LASER encoder to convert bilingual parallel texts into semantic vectors based on (Lewis et al., 2020), thereby constructing a retrieval-augmented context prompting mechanism. For experimental configurations, the temperature parameter was set to 0.7 for both LLMs, with NLLB-200-distilled-600M using a beam size of 10.

As shown in Table 2, the evaluation results on the Flores-200 devtest dataset demonstrate consistent improvements across multiple metrics. The proposed method achieves substantial BLEU score increases of 18.79 for Chinese-Vietnamese (ZH-VI) and 13.55 for Chinese-Malay (ZH-MS) translations compared to baseline systems. While most language pairs demonstrate consistent improvements in COMET scores, Chinese-Burmese (ZH-MY) translations fail to surpass baseline performance, revealing unique challenges in semantic alignment for this particular language combination. These findings validate two key advantages of our approach: the bilingual parallel sentence retrieval mechanism effectively enhances translation quality through precise context matching, while the dynamic retrieval path adaptation improves handling of low-frequency linguistic patterns. The combined strategies enable more accurate capture of grammatical structures and expression patterns in low-resource languages without compromising semantic coherence. The particularly strong results

Table 2: Machine translation results of the Flores-200 **devtest** dataset based on LLMs with parallel sentence pair retrieval tree (PSP-RT).

Models	ZH-VI		ZH-MY		ZH-ID		ZH-MS	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
NLLB-Distilled	28.68	78.19	22.38	**71.93**	16.33	77.12	13.80	74.86
Qwen 2.5 7B	41.17	75.02	8.53	44.58	17.47	79.76	16.02	70.31
+5-Shots	32.82	81.54	12.52	48.51	16.88	83.55	15.98	78.10
+Navie RAG	41.82	77.56	14.21	45.43	18.71	82.92	16.23	79.39
+Our Method	40.39	**81.64**	**32.02**	51.17	18.60	81.39	16.73	79.93
Llama 3.1 8B	41.38	70.71	18.14	45.46	10.58	79.37	26.23	73.37
+5-Shots	44.90	81.46	23.95	53.51	**19.73**	**84.79**	16.41	**80.85**
+Navie RAG	43.22	71.63	19.12	44.39	18.65	82.19	21.93	79.34
+Our Method	**47.44**	79.46	24.06	55.91	17.79	81.89	**27.35**	79.68

Table 3: Ablation Study of Key Pipeline Components. Reranker module removal and embedding substitution effects on translation quality.

Models	ZH-VI		ZH-MY		ZH-ID		ZH-MS	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Qwen 2.5 7B								
+ w/o Reranker	49.32	78.96	20.14	52.62	27.84	83.73	22.46	77.10
+ w/ BGE-M3	51.02	82.23	12.39	50.95	13.16	84.62	21.17	78.81
Llama 3.1 8B								
+ w/o Reranker	45.28	77.92	16.22	56.29	24.14	83.32	32.54	79.75
+ w/ BGE-M3	42.98	81.54	14.25	55.58	14.35	84.39	19.78	80.34

on Vietnamese and Malay translations, despite their differing language families, suggest the method's generalizability across typologically distinct low-resource languages.

4.3 Ablation Experiment

To comprehensively validate the effectiveness of each component in our proposed method, we systematically design ablation experiments analyzing three key dimensions: translation pipeline architecture, model scalability, and context configuration. All experiments are conducted on the development sets of respective language pairs from the Flores-200 dataset.

Pipeline Component Analysis. We perform component-wise ablation studies on the translation pipeline. The removal of our task-fine-tuned reranker model leads to measurable degradation in translation quality, demonstrating its critical role in filtering low-relevance candidates while preserving semantically optimal matches from the retrieval results. Similarly, substituting the original LASER embeddings with BGE-M3(Chen et al., 2024a) results in performance deterioration, revealing fundamental differences in cross-lingual representation efficacy. While BGE-M3 excels in general semantic tasks, its inferior performance compared to the MT-optimized LASER model underscores the importance of task-specific embedding architectures for low-resource translation scenarios. Results are presented in Table 3.

Parameter-Scale Generalization. We evaluate our method's scalability using Gemma2-27B and Llama3.3-70B models(Team et al., 2024; Touvron et al., 2023). As shown in Figure 3, the proposed method consistently outperforms conventional zero-shot and five-shot baselines across both model scales, demonstrating stable performance improvements regardless of parameter count. This stability suggests that our core architecture effectively circumvents the diminishing returns typically associated with mere model scaling, instead leveraging optimized context utilization to enhance the models' inherent translation capabilities.This

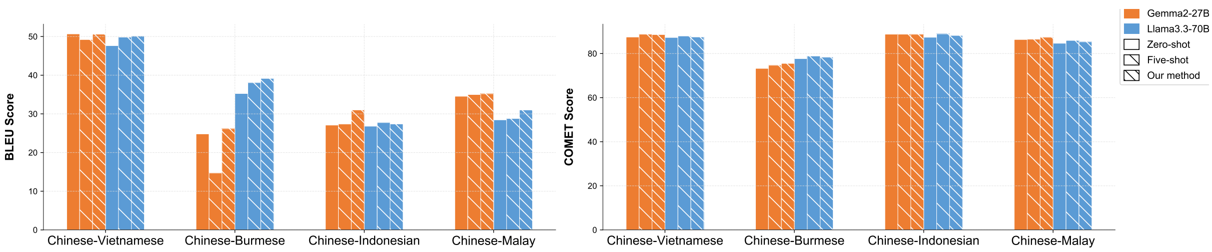

Figure 3: Translation Performance Across Model Scales.(Gemma2-27B and Llama3.3-70B).

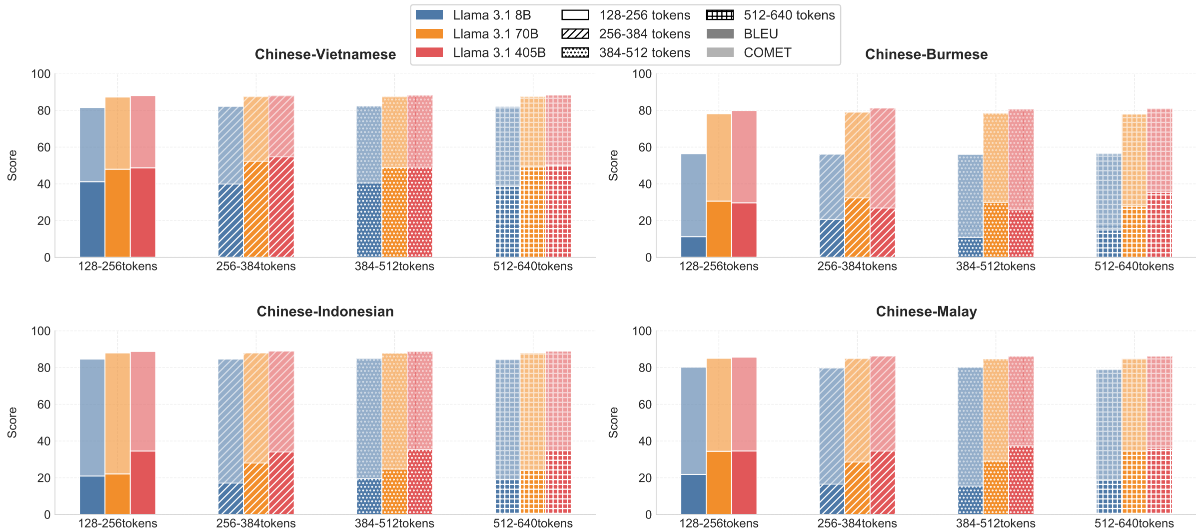

Figure 4: Low-Resource Machine Translation Performance by Context Length and Language Pair on Llama 3.1 Series.

architecture-agnostic effectiveness confirms the method's practical utility for diverse deployment environments.

Prompt Length Sensitivity. We conduct a systematic evaluation of context length effects using Llama3.1 models at three distinct scales (8B, 70B, and 405B parameters), testing window sizes from 128 to 640 tokens with 128-token increments (Fig.4). Our method shows consistent improvements across all model sizes, with distinct optimization patterns: smaller models (8B) perform best with shorter contexts (128-256 tokens), medium models (70B) benefit from extended windows (up to 512 tokens), while the largest model (405B) maintains stable quality across all lengths. The consistent performance improvements across all model sizes further validate our method's effectiveness in enhancing low-resource machine translation through optimized context utilization.

5 Conclusion

This paper presents a low-resource machine translation framework for LLMs that leverages bilingual parallel sentence retrieval. By developing a dynamic semantic retrieval mechanism coupled with context-aware prompt optimization, our approach achieves significant performance improvements in low-resource language scenarios. The framework's core innovation combines the establishment of a cross-lingual semantic retrieval space enabling precise contextual matching of target sentences with the implementation of an adaptive reranking module that simultaneously enhances semantic relevance while effectively eliminating noise. Through comprehensive experiments on multiple Southeast Asian low-resource language pairs, the proposed method demonstrates substantial gains in translation quality.

Limitations

This study primarily targets specific language pairs including Chinese-Vietnamese and Chinese-Burmese, while its applicability to other language families and high-resource scenarios requires further investigation. The real-time retrieval mechanism encounters computational efficiency challenges when processing lengthy texts, particularly increased retrieval latency and memory consumption, which necessitates optimization for improved response speeds in practical applications. Future work will focus on optimizing retrieval efficiency, developing lightweight deployment solutions, and exploring cross-modal knowledge transfer for low-resource translation to further enhance machine translation performance through multimodal knowledge fusion.

Ethics Statement

This study utilizes publicly available datasets and follows standard research protocols for machine translation. All experiments are conducted using open-source models without modification to their original architectures. The research does not involve human subjects or sensitive data, and focuses solely on improving translation quality for low-resource languages. Potential biases in the source datasets may propagate to translation outputs, which should be considered in real-world applications.

References

Firoj Alam, Shammur Absar Chowdhury, Sabri Boughorbel, and Maram Hasanain. 2024. Llms for low resource languages in multilingual, multimodal and dialectal settings. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33.

Mikel Artetxe and Holger Schwenk. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings. *arXiv preprint arXiv:1811.01136*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Artem Babenko and Victor Lempitsky. 2014. The inverted multi-index. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1247–1260.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024b. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation. In *proceedings of the 52nd annual meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Zhangchi Feng, Dongdong Kuang, Zhongyuan Wang, Zhijie Nie, Yaowei Zheng, and Richong Zhang. 2024. Easyrag: Efficient retrieval-augmented generation framework for automated network operations. *arXiv preprint arXiv:2410.10315*.

Yunfan Gao, Yun Xiong, Meng Wang, and Haofen Wang. 2024. Modular rag: Transforming rag systems into lego-like reconfigurable frameworks. *arXiv preprint arXiv:2407.21059*.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yizheng Huang and Jimmy Huang. 2024. A survey on retrieval-augmented text generation for large language models. *arXiv preprint arXiv:2404.10981*.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2019. Filtered pseudo-parallel corpus improves low-resource neural machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(2):1–16.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Langauge Technology (HLT-NAACL 2003)*, pages 48–54. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and 1 others. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023. Making large language models a better foundation for dense retrieval. *Preprint*, arXiv:2312.15503.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual meeting of the Association for Computational Linguistics*, pages 295–302.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. *arXiv preprint arXiv:2012.15674*.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Ratish Puduppully, Anoop Kunchukuttan, Raj Dabre, Ai Ti Aw, and Nancy F Chen. 2023. Decomposed prompting for machine translation between related languages using large language models. *arXiv preprint arXiv:2305.13085*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.(2018).

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

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

Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Kunal Sawarkar, Abhilasha Mangal, and Shivam Raj Solanki. 2024. Blended rag: Improving rag (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 155–161. IEEE.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Introducing the asian language treebank (alt). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1574–1578.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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.

Xintao Wang, Qianwen Yang, Yongting Qiu, Jiaqing Liang, Qianyu He, Zhouhong Gu, Yanghua Xiao, and Wei Wang. 2023. Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases. *arXiv preprint arXiv:2308.11761*.

Zilong Wang, Zifeng Wang, Long Le, Huaixiu Steven Zheng, Swaroop Mishra, Vincent Perot, Yuwei Zhang, Anush Mattapalli, Ankur Taly, Jingbo Shang, and 1 others. 2024. Speculative rag: Enhancing retrieval augmented generation through drafting. *arXiv preprint arXiv:2407.08223*.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. *arXiv preprint arXiv:2109.07684*.

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2024. Unsupervised information refinement training of large language models for retrieval-augmented generation. *arXiv preprint arXiv:2402.18150*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Soyoung Yoon, Eunbi Choi, Jiyeon Kim, Hyeongu Yun, Yireun Kim, and Seung-won Hwang. 2024. Listt5: Listwise reranking with fusion-in-decoder improves zero-shot retrieval. *arXiv preprint arXiv:2402.15838*.

Kun Zhu, Xiaocheng Feng, Xiyuan Du, Yuxuan Gu, Weijiang Yu, Haotian Wang, Qianglong Chen, Zheng Chu, Jingchang Chen, and Bing Qin. 2024a. An information bottleneck perspective for effective noise filtering on retrieval-augmented generation. *arXiv preprint arXiv:2406.01549*.

Shaolin Zhu, Menglong Cui, and Deyi Xiong. 2024b. Towards robust in-context learning for machine translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16619–16629.

A Translation Prompts.

Table 4: Translation Prompt Strategies.

Scenario	**Translation Prompts**
Zero-shot	Translate the Chinese into [target language]. Do not output any hints or explanations other than the results. `Translate: [input]`

Five-shot	Translate the final Chinese into [target language] according to the provided prompts. Do not output any hints or explanations other than the results.
	`Prompts:`
	`[shot 1 reference]`
	`[shot 2 reference]`
	`...`
	`Translate: [input]`

Our Method	You are a professional Chinese to [target language] translator. Please strictly abide by:
	1. Reference Prompts for Translation.
	2. Output only the translation results without any explanation.
	`Prompts:`
	`[shot 1 reference]`
	`[shot 2 reference]`
	`...`
	`Translate: [input]`

Table 4 details the core translation prompt templates used in our experiments. The zero-shot baseline employs straightforward translation instructions requiring only target language output without examples. The five-shot baseline augments this with five bilingual demonstration pairs to enable

Table 5: Machine translation results of the Flores-200 **dev** dataset based on LLMs with parallel sentence pair retrieval tree (PSP-RT).

Models	ZH-VI		ZH-MY		ZH-ID		ZH-MS	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
NLLB-200-Distilled	43.08	79.04	14.04	**73.18**	21.04	77.52	**37.93**	75.10
Qwen 2.5 7B	45.40	78.16	17.17	44.71	10.69	80.68	17.42	70.02
+5-Shots	47.23	**81.06**	14.74	48.22	15.12	84.09	14.96	77.61
+Navie RAG	46.78	78.02	21.97	47.59	19.26	83.77	20.11	77.58
+Our Method	**52.95**	79.67	**22.39**	51.58	**30.04**	**84.65**	21.51	78.87
Llama 3.1 8B	42.28	71.61	9.69	45.90	16.77	78.55	19.62	72.44
+5-Shots	44.13	78.62	14.96	52.41	27.26	84.62	20.47	**80.51**
+Navie RAG	40.08	69.93	13.97	48.64	25.78	83.97	25.67	79.43
+Our Method	44.56	76.50	18.19	65.06	26.43	84.34	34.83	80.20

few-shot learning. Our method introduces a professional translator role declaration to strengthen behavioral constraints, while incorporating dynamically retrieved bilingual examples as contextual references. All templates strictly limit outputs to translation content without additional explanations.

B Experimental Results On Flores-200 dev Datasets.

The evaluation results on the Flores-200 dev dataset demonstrate consistent performance improvements across multiple language pairs, as detailed in Table 5. The proposed method achieves significant BLEU score increases of 9.87 for Chinese-Vietnamese (ZH-VI) alongside a 0.63 gain in COMET score, indicating substantial improvements in both lexical and semantic translation quality. Similar improvements emerge for Chinese-Burmese (ZH-MY) and Chinese-Indonesian (ZH-ID), which show BLEU gains of 8.35 and 9.00 respectively, confirming the method's effectiveness across diverse Southeast Asian languages. While Chinese-Malay (ZH-MS) translations show slightly more modest results, remaining 3.10 BLEU points below the baseline, the overall pattern reveals robust performance gains that validate our approach's ability to handle varying linguistic characteristics.