Don’t Take it Literally!
Idiom-aware Vietnamese Translation via In-context Learning

Luan Thanh Nguyen[1,2], **Parisa Kordjamshidi**[3]
[1]Faculty of Information Science and Engineering, University of Information Technology,
Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]Department of Computer Science and Engineering, Michigan State University
luannt@uit.edu.vn, kordjams@msu.edu

Abstract

The translation of idiomatic expressions often results in misunderstandings and inaccuracies, affecting everyday communication as well as machine translation systems. This paper introduces **Idi**om-**a**ware Vietnamese **T**ranslation (IDIAT), a new framework for the evaluation of idiomatic translation for Vietnamese, along with state-of-the-art results for this task. We collect and curate a high-quality Vietnamese-English idiom set that serves as a resource for in-context learning (ICL). IDIAT’s evaluation benchmark includes both idiomatic and non-idiomatic text pairs to assess general translation quality and idiomatic translation performance. We leverage ICL in large language models to augment few-shot demonstrations with idiom and topic descriptions and consequently improve the translation accuracy. Empirical results demonstrate that our IDIAT-based ICL outperforms traditional supervised methods using only a few data samples. Multiple evaluations confirm the effectiveness of our proposed approach. Though focusing on the Vietnamese language, our approach advances idiomatic translation and contributes to the development of culturally aware translation systems, paving the way for future research in low-resource languages. The experimental materials are publicly available[1].

1 Introduction

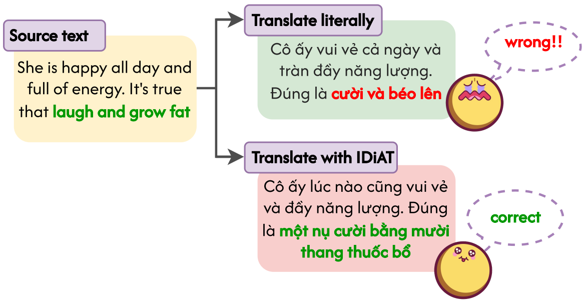

Source text
She is happy all day and full of energy. It's true that laugh and grow fat
Translate literally
Cô ấy vui vẻ cả ngày và tràn đầy năng lượng. Đúng là cười và béo lên
wrong!!
Translate with IDiAT
Cô ấy lúc nào cũng vui vẻ và đầy năng lượng. Đúng là một nụ cười bằng mười thang thuốc bổ
correct

Figure 1: While the literal translation of the idiom "laugh and grow fat" produces an incorrect and unnatural result in Vietnamese, the IDIAT framework captures the idiomatic meaning, yielding a culturally appropriate and accurate translation.[2]

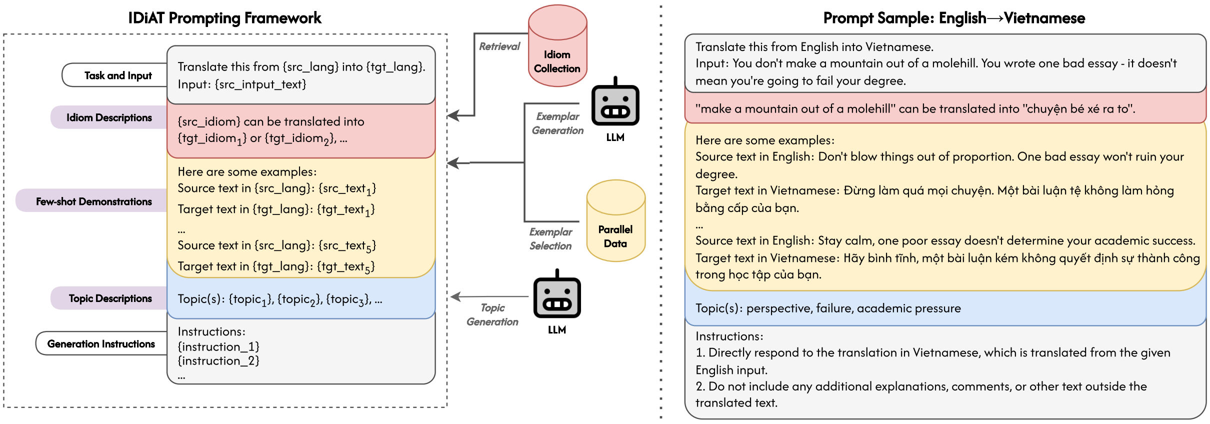
Idiomatic expressions pose a significant challenge in real-life conversation and machine translation models (Ahmed and Saadoun, 2024; Vula and TyfekÃ, 2024). These expressions often carry meanings that are not directly translatable, leading to potential misunderstandings and inaccuracies. In the context of neural machine translation, idioms can result in translations that are either overly literal or miss the intended meaning entirely, thereby compromising the quality and fluency of the output (Aldelaa et al., 2024). Figure 1 depicts the contrast between the shortcomings of literal translation and the effectiveness of idiom-aware translation.

Recent advancements in large language models (LLMs) have shown promise in addressing these challenges. LLMs possess remarkable disambiguation and contextual understanding abilities, allowing them to generate translations more aligned with human expectations (Xu et al., 2024; Zhang et al., 2023). Following that, the emergence of ICL has transformed how language models approach tasks by allowing them to learn from examples provided within the input prompt, eliminating the need for task-specific fine-tuning (Brown et al., 2020; Gao et al., 2021). This general adaptability has shown particular promise in addressing linguistic ambiguity and enabling idiomatic translation, where few-shot prompting helps models infer context-specific meanings. For specific tasks such as translation, the ability of ICL, which captures subtle language features, is especially valuable and can potentially enhance the generation performance.

Vietnamese is a tonal and analytic language characterized by its rich vocabulary and complex syntactic structures, reflecting the region’s cultural and historical depth (Francis, 2023; Jamieson, 2023;

[1] https://github.com/tarudesu/IDiAT

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1795–1814
December 20-24, 2025 ©2025 Association for Computational Linguistics



Translate this from {src_lang} into {tgt_lang}.
Input: {src_intput_text}

{src_idiom} can be translated into {tgt_idiom$_1$} or {tgt_idiom$_2$}, ...

Here are some examples:
Source text in {src_lang}: {src_text$_1$}
Target text in {tgt_lang}: {tgt_text$_1$}
...
Source text in {src_lang}: {src_text$_5$}
Target text in {tgt_lang}: {tgt_text$_5$}

Topic(s): {topic$_1$}, {topic$_2$}, {topic$_3$}, ...

Instructions:
{instruction_1}
{instruction_2}
...

Translate this from English into Vietnamese.
Input: You don't make a mountain out of a molehill. You wrote one bad essay - it doesn't mean you're going to fail your degree.

"make a mountain out of a molehill" can be translated into "chuyện bé xé ra to".

Here are some examples:
Source text in English: Don't blow things out of proportion. One bad essay won't ruin your degree.
Target text in Vietnamese: Đừng làm quá mọi chuyện. Một bài luận tệ không làm hỏng bằng cấp của bạn.
...
Source text in English: Stay calm, one poor essay doesn't determine your academic success.
Target text in Vietnamese: Hãy bình tĩnh, một bài luận kém không quyết định sự thành công trong học tập của bạn.

Topic(s): perspective, failure, academic pressure

Instructions:
1. Directly respond to the translation in Vietnamese, which is translated from the given English input.
2. Do not include any additional explanations, comments, or other text outside the translated text.

Figure 2: The IDIAT Prompting Framework consists of five key components: (1) *Task and Input*, which defines the task and input for the LLM; (2) *Few-shot Demonstrations*, providing exemplar translations to guide the model; (3) *Idiom Descriptions*, offering idiomatic translations for nuanced understanding; (4) *Topic Descriptions*, outlining contextual related topics; and (5) *Generation Instructions*, detailing formatting instructions for output generation.

Tran, 2024). Among its linguistic features, idioms are significant, often conveying figurative meanings that extend beyond their literal interpretations (Giang, 2023a,b; Hanh et al., 2023). Consequently, translating these expressions based on their contextual and cultural significance is crucial to achieving accurate and culturally resonant translations. Nonetheless, existing translation approaches often fail to adequately address these rich linguistic features, frequently prioritizing literal translations over capturing the deeper cultural and contextual nuances in the language.

To tackle the challenge of idiom translation in low-resource languages like Vietnamese, we propose a framework with a novel evaluation resource called IDIAT. While our new resource makes the evaluation of Vi↔En for idiom-aware translation possible, our proposed idiom-aware-ICL harnesses the power of LLMs to convey the meanings of idioms in the target language accurately.

In our best idiom-aware-ICL practice, we used three key components of few-shot demonstrations, idiom descriptions, and topic descriptions. These components enhanced translation performance, particularly for idiomatic expressions. By incorporating contextual information and relevant examples, we improved both the accuracy and fluency of translations, addressing the shortcomings of traditional methods that often overlook the nuances of idiomatic language. The contributions of this work are summarized in three main key points:

- We create a new evaluation benchmark IDIAT for idiom-aware Vi↔En translation, including a high-quality bilingual idiom collection;
- We propose an IDIAT-based ICL pipeline that leverages the strengths of ICL to enhance idiomatic translation for Vietnamese;
- We present an extensive experimental study using our curated resource alongside existing translation datasets, thoroughly demonstrating the effectiveness of our IDIAT-based ICL pipeline across diverse evaluation metrics.

Source	**Have idiom**	**No idiom**
PhoMT (Doan et al., 2021)	181	664
Official dictionary (Lã, 1995)	155	0
Total	336	664

Table 1: The distribution of 1,000 instances in the IDIAT benchmark evaluation test set.

2 Data Creation

2.1 IDiAT Benchmark

Recognizing the lack of idiomatic expressions in existing Vi↔En translation benchmarks, we construct a high-quality benchmark to assess both general and idiomatic translation. We start by filtering the test split of the PhoMT dataset (Doan et al., 2021) to extract idiom-containing samples, then add non-idiomatic examples from PhoMT to support general translation evaluation. To further expand coverage, we include entries from the official Vietnamese-English idiom dictionary (Lã, 1995). The final evaluation set contains 1,000 samples, with their distribution shown in Table 1.

2.2 Idiom Collection

Prior work, such as IdiomKB (Ghazvininejad et al., 2023), shows that using context and idiom descriptions in prompts improves idiom understanding. Building on this, we create a large collection of Vietnamese idioms paired with English equivalents to support idiomatic translation via ICL. These bilingual pairs are drawn from an official Vietnamese-English idiom textbook (Nguyen, 2014) and are manually curated to ensure semantic alignment. The final dataset includes 5,000 idiom pairs, providing a valuable resource for both evaluation and research on idiomatic translation in low-resource language settings.

3 IDiAT: Idiom-aware Translation

In this study, we propose IDIAT framework, an effective ICL pipeline for Vi↔En translation, in order to enhance translation performance and its ability to translate idiomatic expressions by integrating various components that provide contextual understanding and guidance for the translation process. Figure 2 illustrates our entire framework.

3.1 Few-shot Demonstrations

The term few-shot demonstrations is recognized as a crucial component of the prompt, guiding LLMs to generate accurate outputs. Moreover, various exemplar selection techniques can impact the performance of LLMs (Wang et al., 2023; Wei et al., 2022b; Chu et al., 2024; Gupta et al., 2023; Kumar et al., 2023; Ye et al., 2023; Liu et al., 2024). This study investigates various exemplar selection methods to enhance few-shot prompting for LLMs, which are crucial for guiding accurate model outputs. These methods include *Random Sampling*, a simple yet quality-variable approach; *SBERT Similarity Ranking*, which selects examples based on semantic similarity using Sentence Transformers; and *BM25 Ranking*, which retrieves contextually relevant examples through probabilistic scoring. Additionally, the study explores *LLM-generated Demonstrations*, which prompts LLMs to produce its own examples, leveraging its internal reasoning to create context-aware and idiomatic translations.

3.2 Idiom Descriptions

Using dictionaries as references (Lu et al., 2024) for prompting has proven effective in enhancing the performance of LLMs in translation tasks. Specifically, including idiom descriptions has shown potential in improving idiomatic translation and context disambiguation (Li et al., 2024). In this research, we implement two approaches: collection-based idiom retrieval from a curated collection and using LLMs as generators for idiom meanings to leverage ICL for enhancing translation.

First, the collection-based method incorporates three retrieval techniques: *1) Exact Matching*, which retrieves idioms that precisely match the input idiom to ensure equivalence; *2) Fuzzy Matching* with a threshold, which retrieves similar but not identical idioms using a similarity threshold, making it effective for handling idiom variants; and *3) BM25 Ranking*, which ranks idioms based on their relevance to the input idiom to retrieve contextually appropriate equivalents.[3]

On the target language side, since an idiom may have multiple equivalent expressions, we adopt two strategies to incorporate these into the translation prompt: 1) *Use all* matching idioms from the collection, or 2) *Use Top-1* matching based on cross-lingual similarity scores computed with a multilingual Sentence Transformer (Reimers and Gurevych, 2020).

For the idiom description, we prompt the model to produce either the equivalent idiom in the target language or its literal translation if no direct equivalent exists. This approach assesses the LLM's ability to understand idiomatic expressions, particularly in low-resource languages like Vietnamese.

3.3 Topic Descriptions

He et al. (2024) demonstrated the effectiveness of using topic descriptions in prompting to enhance translation task performance. This approach outlines the contextual topics relevant to the task, aiding the model in maintaining coherence and relevance in its output. By incorporating this component, the translations better align with the intended meaning, thereby improving the overall performance of LLMs in translation.

4 Experimental Results

In this section, we outline the experimental settings used to evaluate the performance of our IDIAT-based ICL pipeline on the curated benchmark, in the context of idiomatic translation.

Model. We primarily present experimental results on the commercial LLM GPT-4o-mini, a compact

[3] We set the threshold for Fuzzy Matching and BM25 Ranking at 0.7 and 0.3, respectively.

variant of GPT-4o (OpenAI et al., 2024). Additionally, we evaluate several open-source LLMs, including Qwen (Yang et al., 2024), LLaMA (Grattafiori et al., 2024), and Gemma (Team et al., 2024) (see Section 5.3).

Data. All experiments and evaluations are conducted on the IDIAT benchmark test set and the curated Vi-En idiom collection (see Section 2).

SOTA. The current state-of-the-art for Vi↔En translation is represented by the EnViT5-translation[4] model (Ngo et al., 2022), which has been fine-tuned on 4M+ Vi-En parallel pairs. This model serves as a benchmark for evaluating the performance of our proposed methods.

Baseline. We use zero-shot prompting to evaluate performance without fine-tuning or in-context examples, enabling a clear comparison.

4.1 Evaluation Metrics

Automated Metrics. To assess the translation performance, we utilize two key metrics: sacreBLEU[5] (Post, 2018) and COMET[6] (Rei et al., 2020). While sacreBLEU focuses on measuring n-gram overlap between the predictions and references, offering a standard method for evaluating translation quality, COMET provides a deeper assessment of semantic alignment, making it particularly effective for capturing the nuances of idiomatic expressions.

LLM-based Metric. Utilizing LLMs for assessing the idiomatic translation quality across different language pairs has recently shown their benefits (Li et al., 2024). In this study, we report the GPT-score using the GPT-4o model as an evaluator on the IDIAT evaluation benchmark dataset[7].

Human-based Metric. To ensure comprehensive evaluation, we also conduct human evaluations to assess the translations. Each annotator is provided with detailed annotation guidelines, illustrated in Appendix H, and asked to select the best translation among three approaches (SOTA, Baseline, and IDIAT). The final evaluation results are averaged to provide a robust measure of translation quality.

4.2 Results

Table 2 summarizes our findings. We selected the best ICL method in IDIAT per translation direction based on the highest COMET score. The optimal integration is BM25 Ranking (Few-shot, En→Vi) or LLM Generation (Few-shot, Vi→En) + Use-all with Fuzzy Matching (Idiom) + (Topic).

IDIAT outperforms the baseline in all tests and both directions. The proposed framework, IDIAT, consistently performs better than the baseline zero-shot prompting method across all evaluation metrics. For instance, in the En→Vi direction, IDIAT achieves a BLEU score of 35.13 and a COMET score of 57.38, compared to the baseline scores of 32.98 and 54.51, respectively. Similarly, in the Vi→En direction, IDIAT scores 33.81 (BLEU) and 60.64 (COMET), significantly surpassing the baseline scores of 29.88 and 52.90. These results highlight the effectiveness of the IDIAT-based ICL framework, compared to those of the baseline, in enhancing idiomatic translation quality.

Idiom descriptions benefit LLMs in idiomatic translation. The experimental results clearly demonstrate that including idiom descriptions significantly enhances the performance of the translation model for idiomatic expressions. When examining the performance on instances that contain idioms, we observe that all methods utilizing idiom descriptions yield improved results in both translation directions. For instance, the BLEU score for idioms in the En→Vi direction increases to 31.40 with IDIAT, compared to 27.71 for the SOTA model, indicating a substantial improvement. Similarly, in the Vi→En direction, the BLEU score for idioms rises to 32.29, surpassing the SOTA score.

Moreover, the COMET scores also reflect substantial gains. In the En→Vi direction, the COMET score reaches 52.90 with IDIAT, compared to 32.12 (SOTA), indicating a more substantial alignment with human evaluators' expectations. In the Vi→En direction, the COMET score for idioms improves to 32.29, exceeding the SOTA performance.

Despite the variability of using LLM-generated idiom descriptions, it still benefits the translation performance. The BLEU score for the LLM-generated approach reaches 27.63 in the Vi→En direction, which is higher than the baseline zero-shot prompting score of 25.29. This consistent improvement across all methods suggests that idiom descriptions provide critical contextual information that aids the model in understanding and accurately translating idiomatic expressions, which are often nuanced and context-dependent.

LLMs show their effectiveness in generating human-like translation. The COMET scores for all cases of using the LLM across all methods

[4] https://huggingface.co/VietAI/envit5-translation

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

[6] Unbabel/wmt20-comet-da

[7] We re-implement Li et al. (2024)'s prompt.

Methods		En→Vi						Vi→En					
		All		✓ idioms		✗ idioms		All		✓ idioms		✗ idioms	
		BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
SOTA: Supervised Fine-tuning Sequence-to-Sequence Models													
EnViT5-base		36.76	50.08	27.71	32.12	39.86	59.17	32.58	48.01	25.50	31.55	35.18	56.33
Baseline: Zero-shot Prompting with LLMs													
Zero-shot Prompting		32.98	54.51	25.75	44.93	35.46	59.36	29.88	52.90	25.29	40.49	32.57	59.18
Proposed Methods: In-context Learning with LLMs													
Component 1: Few-shot Demonstrations													
Random Sampling		33.88	54.39	26.79	44.86	36.30	59.21	29.85	52.98	25.44	41.09	31.46	59.00
SBERT Ranking		33.54	54.30	26.51	44.94	35.97	59.04	30.02	52.85	25.48	39.98	31.67	59.36
BM25 Ranking		**33.88**	**54.52**	**26.84**	**45.09**	**36.30**	**59.30**	29.93	52.75	25.41	40.15	31.57	59.12
LLM Generation		31.00	53.03	24.51	43.89	33.30	57.66	**32.35**	**58.11**	**27.63**	**43.78**	**34.07**	**65.36**
Component 2: Idiom Descriptions													
Use all retrieved idioms	Exact Matching	34.31	57.00	30.96	52.36	N/A		31.27	54.99	30.48	46.72	N/A	
	Fuzzy Matching	34.35	**57.08**	31.11	**52.57**			**31.27**	**55.05**	**30.49**	**46.88**		
	BM25 Ranking	34.34	56.99	31.06	52.30			31.27	54.96	30.48	46.61		
Use Top-1	Exact Matching	**34.43**	56.67	**31.40**	51.36			31.16	54.80	30.07	46.15		
	Fuzzy Matching	34.40	56.69	31.30	51.41			31.16	54.81	30.07	46.16		
	BM25 Ranking	34.40	56.72	31.26	51.51			31.12	54.78	30.07	46.32		
LLM Generation		33.23	53.28	26.59	41.26			30.44	53.57	27.34	42.49		
Component 3: Topic Description													
LLM Generation		33.77	55.10	26.65	46.17	36.22	59.62	29.67	53.31	25.17	41.73	31.32	59.17
IDIAT (with best retrieval approaches)		**35.13**	**57.38**	**31.40**	**52.90**	**36.41**	**59.65**	**33.81**	**60.64**	**32.29**	**51.22**	**34.33**	**65.41**

Table 2: The performance is reported using IDIAT test set. Results are shown for all data ("All"), idiom-containing subsets ("✓ idioms"), and non-idiom subsets ("✗ idioms"). Bold values indicate the best-performing method for each component tested across multiple approaches. The bold results for IDIAT highlight its superior performance over the baseline. Metrics include BLEU and COMET (higher is better). All results use GPT-4o-mini. N/A indicates ("✗ idioms") prompts match the baseline since no idioms are included.

Methods	En→Vi						Vi→En					
	All		✓idioms		✗idioms		All		✓idioms		✗idioms	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Baseline	32.98	54.51	25.75	44.93	35.46	59.36	29.88	52.90	25.29	40.49	31.57	59.18
IDIAT	35.13	57.38	31.40	52.90	36.41	59.65	33.81	60.64	32.29	51.22	34.33	65.41
w/o few-shot	35.09 $_{\downarrow 0.04}$	57.70 $_{\uparrow 0.32}$	31.89 $_{\uparrow 0.49}$	54.31 $_{\uparrow 1.41}$	36.17 $_{\downarrow 0.24}$	59.42 $_{\downarrow 0.23}$	31.15 $_{\downarrow 2.66}$	55.60 $_{\downarrow 5.04}$	30.46 $_{\downarrow 1.83}$	47.95 $_{\downarrow 3.27}$	31.41 $_{\downarrow 2.92}$	59.47 $_{\downarrow 5.94}$
w/o idiom	33.89 $_{\downarrow 1.24}$	54.53 $_{\downarrow 2.85}$	26.77 $_{\downarrow 4.63}$	44.48 $_{\downarrow 8.42}$	-	-	32.83 $_{\downarrow 0.98}$	58.30 $_{\downarrow 2.34}$	28.16 $_{\downarrow 4.13}$	44.48 $_{\downarrow 6.74}$	-	-
w/o topic	34.82 $_{\downarrow 0.31}$	57.09 $_{\downarrow 0.29}$	31.18 $_{\downarrow 0.22}$	53.46 $_{\uparrow 0.56}$	36.06 $_{\downarrow 0.35}$	58.93 $_{\downarrow 0.72}$	33.72 $_{\downarrow 0.09}$	60.49 $_{\downarrow 0.15}$	32.32 $_{\uparrow 0.03}$	51.24 $_{\downarrow 0.02}$	34.19 $_{\downarrow 0.14}$	65.16 $_{\downarrow 0.25}$

Table 3: Ablation study results comparing BLEU and COMET scores across En↔Vi idiomatic translation tasks. The study examines the impact of removing individual components from the IDIAT framework - few-shot demonstrations (w/o few-shot), idiom descriptions (w/o idiom), and topic descriptions (w/o topic). Subscript values indicate performance changes relative to the complete IDIAT, with ↓ for decreases and ↑ for improvements.

consistently outperform the SOTA model, indicating that its translations are more accurate and closely aligned with human evaluators' expectations. Specifically, the COMET scores obtained by IDIAT in both En→Vi and Vi→En directions surpass the SOTA by 7.3 and 12.63, respectively. This further suggests that LLMs are capable of producing translations that feel natural and are contextually appropriate, surpassing traditional models in producing human-like quality.

4.3 Ablation Study on Idiomatic Translation

The ablation study in Table 3 highlights the contributions of each IDIAT framework component:

w/o few-shot. Removing few-shot examples slightly lowers BLEU (En→Vi drops from 35.13 to 35.09) but raises COMET (57.38 to 57.70). This suggests that while the few-shot demonstrations contribute positively to overall performance, their absence does not drastically hinder the model's ability to generate idiomatic translations, particularly in terms of semantic alignment. However, the BLEU score for idiomatic instances still slightly increases, indicating that the model can still leverage its learned knowledge effectively even without explicit few-shot examples.

w/o idiom. The removal of idiom descriptions results in a decrease across all metrics, indicating that these descriptions are crucial for maintaining the quality of idiomatic translations. This decline underscores the importance of idiom descriptions in

Methods	GPT-score	
	En→Vi	Vi→En
Topline with EnViT5-base	1.75	1.79
Baseline with Zero-shot Prompting	2.12	2.35
IDiAT (ours)	**2.41**	**2.63**

Table 4: Comparison of GPT-scores for translation across three approaches. Scores are averaged across the 100-sample set, with a scale of 1-3, where higher scores indicate better translation quality.

providing the necessary context for accurate translation, as idioms often carry meanings that are not directly translatable without additional context.

w/o topic. Removing topic descriptions causes slight performance declines, though the En→Vi COMET score increases marginally. This could suggest that while topic descriptions generally help maintain coherence and relevance in translations, the model may still perform adequately in terms of semantic similarity without them.

5 Analysis and Discussion

We further analyzed the results using GPT-score, human evaluation, and translation quality metrics. We also present experimental results for other open-source LLMs and low-resource languages.

5.1 GPT-score

We calculate the GPT-score on 100 idiom-containing samples randomly selected from the IDiAT benchmark dataset for this experiment.

The results in Table 4 show that our proposed method, IDiAT, achieves the highest GPT-scores, surpassing both the Topline and Baseline in both translation directions. By leveraging multiple ICL techniques, IDiAT effectively addresses idiomatic translation challenges, outperforming zero-shot prompting and even traditional supervised fine-tuning on large-scale parallel data. These findings highlight the value of specialized methods and also the relevance of GPT-score in assessing translation quality for idiomatic expressions.

5.2 Human Evaluation

The human evaluation is also conducted on the 100-sample set to assess translation quality. Five undergraduate students are hired for this task[8], and each student is asked to select the best translation from the options provided by three methods: Topline, Baseline, and IDiAT . The evaluation setup, question template for each sample, as well as the guidelines for annotation are in Appendix H.

Table 6 provides the results of the human evaluation, showcasing the performance of the three translation methods as judged by human. IDiAT again outperforms its counterparts, achieving human evaluation scores of 82.4% for En→Vi and 83.0% for Vi→En. These results are markedly higher than those of the Topline (22.8% and 23.6%) and the Baseline (39.8% and 50.2%).

The significant margin by which IDiAT exceeds the other methods demonstrates its ability to produce translations that better align with human preferences, especially for idiomatic expressions. The out-performance is across both directions.

Interestingly, the scores achieved by the Baseline even outperform the Topline, indicating that zero-shot prompting, despite its lack of explicit fine-tuning on parallel data, can leverage the generalization capabilities of LLMs to handle idiomatic expressions more effectively than a supervised model trained on extensive but conventional parallel datasets. This suggests that traditional fine-tuning approaches may struggle with idiomatic translations when the training data lacks sufficient idiomatic coverage, whereas LLMs benefit from the diverse linguistic patterns captured during the pre-training phase of the language model.

5.3 Generalization on Models and Languages

Besides the results achieved by GPT-4o-mini presented in Section 4, we also conduct multiple implementations on other LLMs and other languages.

Robustness Across Open-Source LLMs. We further assess the effectiveness of IDiAT-based ICL pipeline across a range of open-source LLMs of varying sizes, including Qwen2.5, LLaMA-3.1 and 3.2, and Gemma2, spanning from 494M to 7.62B parameters, as detailed in Appendix A. Regardless of model scale, IDiAT consistently improves translation quality in both En→Vi and Vi→En directions. Notably, it leads to substantial gains in translating idiomatic expressions, as evidenced by the improvement margins between the baseline (✗) and IDiAT-enhanced (✓) outputs.

IDiAT with Low-Resource Languages. Beyond the Vi–En pair, we extend our study to X↔English translation tasks, where X includes mid-resource (Japanese, Korean), low-resource (Thai), and ex-

[8] Each student is paid approximately 4 USD for annotating 100 samples, a rate that surpasses the local minimum wage.

Methods	**Translations**	**GPT-score**	**Human**
Vietnamese → English			
Topline	His mom said, "You don't want to run in front of the car, or you're gonna fail your test."	1	✗
Baseline	His mother said, "You shouldn't run with a lantern in front of a car, or you'll fail the exam."	1	✗
IDIAT (ours)	His mother said, "Don't put the cart before the horse, or you might fail the test."	3	✓
Source	Mẹ cậu ấy nói "**Không nên cầm đèn chạy trước ô tô**, nếu không con sẽ thi trượt đấy."		
Reference	"**Don't put the cart before the horse** or you will fail the exam," his mother said.		
English → Vietnamese			
Topline	Ông quyết định chèo xuồng của riêng mình và thành lập công ty riêng.	1	✗
Baseline	Anh ấy quyết định tự chèo thuyền của mình và thành lập công ty riêng.	1	✗
IDIAT (ours)	Anh ấy quyết định tự lực cánh sinh và thành lập công ty riêng của mình.	3	✓
Source	He decided to **paddle his own canoe** and set up his own company.		
Reference	Anh ấy quyết **tự lực cánh sinh** và thành lập công ty của chính mình.		

Table 5: Comparison of generated translations from three methods for Vi↔En idiomatic translation, evaluated by GPT-score and human assessment. Note that ✓ indicates human preference, while ✗ denotes otherwise.

Methods	**Human Evaluation**	
	En→Vi	**Vi→En**
Topline with EnViT5-base	22.8	23.6
Baseline with Zero-shot Prompting	39.8	50.2
IDIAT (ours)	**82.4**	**83.0**

Table 6: Human evaluation scores for three translation approaches. Results are based on pairwise comparisons across the 100-sample set, showing IDIAT achieves significantly higher preference rates in both directions.

tremely low-resource languages (Finnish, Slovenian). The performance improvements, detailed in Appendix B, demonstrate that the IDIAT approach remains effective even in limited-resource settings, consistently enhancing translation quality in both idiomatic and non-idiomatic contexts.

5.4 Qualitative Comparison

Table 5 compares the idiomatic translations of three methods (Topline, Baseline, and IDIAT) for both Vi↔En directions. In the Vi→En, IDIAT accurately translates the idiom "**Không nên cầm đèn chạy trước ô tô**" to "**Don't put the cart before the horse**," while Topline and Baseline provide literal, incorrect translations. Similarly, in the En→Vi, IDIAT translates "**paddle his own canoe**" as "**tự lực cánh sinh**," aligning with the idiomatic meaning, while the other methods offer literal translations. These results highlight IDIAT's effectiveness in handling idioms with cultural and linguistic accuracy, thanks to ICL and idiom-specific fine-tuning. These examples emphasize the ability of IDIAT to identify and generate contextually appropriate idiomatic translations, bridging cultural and linguistic nuances that are often missed by conventional approaches. This success is attributed to the ICL strategies and idiom-specific fine-tuning incorporated in IDIAT, which enable it to go beyond literal translations and achieve human-like fluency in handling idiomatic expressions.

5.5 Impact of Idiom Complexity

To further understand the translation performance across different idiom types, we conducted an in-depth analysis presented in Appendix C and D. This analysis examines idioms along three dimensions: *semantic opacity*, *usage frequency*, and *cultural-linguistic equivalence*. The results show that IDIAT consistently outperforms the zero-shot baseline, particularly with *opaque*, *rare*, and *culturally nuanced idioms*, where literal or semantically mismatched translations often occur. Moreover, in handling *unseen idioms*, IDIAT demonstrates stronger contextual reasoning and idiomatic substitution, resulting in better results in performing translation.

6 Related Work

Recent advancements with the emergence of LLMs and ICL techniques, which led to significant progress in translation and idiomatic expression handling, are reviewed in this section.

6.1 LLMs and ICL in Translation

LLMs, such as the GPT series (Moslem et al., 2023; He et al., 2024; Pang et al., 2025), have revolutionized translation by leveraging pre-trained knowledge from diverse text corpora to generate coherent and contextually appropriate outputs. Their ability to perform few-shot and zero-shot learning enables effective adaptation to low-resource languages, addressing data scarcity challenges while enhancing multilingual proficiency (Babaali et al., 2024; Guo et al., 2024; Merx et al., 2024). A key phenomenon within LLMs that amplifies their effectiveness is in-context learning, which allows them to generalize from examples provided in the input without requiring explicit fine-tuning (Brown et al., 2020; Wei et al., 2022a; Liu et al., 2023). Through ICL, LLMs can dynamically adapt to linguistic variations, improving disambiguation and translation quality across different contexts (Gao et al., 2021; Iyer et al., 2023). By integrating contextual cues and leveraging prior knowledge, LLMs equipped with ICL enhance both the accuracy and cultural appropriateness of translations, making them especially powerful for low-resource languages (Agrawal et al., 2023; Cahyawijaya et al., 2024; Dwivedi et al., 2024).

6.2 Idiomatic Translation Disambiguation

Translating idiomatic expressions presents a significant challenge due to their non-compositional and culturally specific nature. Recent studies have explored the use of LLMs to address this issue. Donthi et al. (2025) introduced two methods: Semantic Idiom Alignment (SIA), which employs pre-trained sentence embeddings to identify semantically similar idioms in the target language, and Language-Model-based Idiom Alignment (LIA), which prompts an LLM to suggest appropriate idiomatic counterparts. Their findings indicate that SIA more effectively preserves idiomatic style across languages such as Chinese, Urdu, and Hindi. Similarly, Castaldo and Monti (2024) examined the impact of prompt design on idiomatic translation quality between English and Italian, revealing that carefully crafted prompts can significantly enhance translation outcomes. Additionally, Li et al. (2024) developed IdiomKB, a multilingual idiom knowledge base constructed using LLMs. IdiomKB provides figurative meanings of idioms, aiding smaller models in achieving more accurate translations. Their approach emphasizes context awareness and scalability, contributing to improved idiomatic translation performance. Collectively, these studies demonstrate the potential of LLMs and associated techniques in improving the cultural and contextual accuracy of idiomatic translation.

6.3 Vietnamese Translation Approaches

Conventional approaches to Vietnamese translation have primarily relied on neural machine translation models (Doan et al., 2021; Minh et al., 2021; Ngo et al., 2022; Pham et al., 2023), which require a large amount of parallel data for training. Building on this foundation, the use of LLMs in translation has emerged with outstanding performance, as demonstrated by projects like DocTranslate[9], which currently achieves state-of-the-art results on the PhoMT dataset. However, this tool is primarily commercial and not publicly available for the research community. Furthermore, to the best of our knowledge, no prior research has specifically addressed the Vietnamese idiomatic translation.

7 Conclusion

This work has explored the potential of in-context learning to enhance idiomatic translation between Vietnamese and English for disambiguation and contextual understanding. Our proposed idiom-based ICL pipeline, called IDIAT, integrates idiom descriptions and relevant topic descriptions in the context and improves the LLMs to generate semantically and culturally relevant translations. This research leverages the strengths of LLMs and ICL to create a robust framework for addressing idiomatic complexities, paving the way for future research. The IDIAT framework can be applied to other low-resource and highly low-resource languages for a more inclusive and effective translation systems that bridge linguistic and cultural gaps.

Limitations

This study has some limitations. First, the experiments were conducted using small and medium-sized LLMs; larger models, with their increased capacity, may achieve better performance and more nuanced translations. Furthermore, the collection of Vietnamese-English idioms used in this study may not be comprehensive, which could affect the model's accuracy in translating idiomatic expressions. Addressing these limitations in future research will enhance the effectiveness and applicability of the IDIAT-based ICL framework across broader contexts and languages.

[9] https://github.com/doctranslate-io/viet-translation-llm

Acknowledgement

The authors acknowledge partial support from USAID Partnership for Higher Education Reform project for work on this manuscript. This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

Thanks to Michigan State University and the Heterogeneous Learning and Reasoning Lab for hosting and supporting the visiting scholar stay.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Saif Saadoon Ahmed and Saif Saadoun. 2024. Translation and semantics: Challenges and strategies in translating english idioms. *Journal of Language Studies. Vol*, 8(3):347–335.

Abdullah S Aldelaa et al. 2024. Investigating problems related to the translation of idiomatic expressions in the arabic novels using neural machine translation. *Theory and Practice in Language Studies*, 14(1):71–78.

Baligh Babaali, Mohammed Salem, and Nawaf R Alharbe. 2024. Breaking language barriers with chatgpt: enhancing low-resource machine translation between algerian arabic and msa. *International Journal of Information Technology*, pages 1–10.

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.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Antonio Castaldo and Johanna Monti. 2024. Prompting large language models for idiomatic translation. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 32–39, Sheffield, United Kingdom. European Association for Machine Translation.

Zheng Chu et al. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1173–1203, Bangkok, Thailand. Association for Computational Linguistics.

Nguyen Giang Dang. 2011. Idiom variants and synonymous idioms in english and vietnamese: The similarities and differences. *VNU Journal of Foreign Studies*, 27(4).

Long Doan, Linh The Nguyen, Nguyen Luong Tran, Thai Hoang, and Dat Quoc Nguyen. 2021. PhoMT: A high-quality and large-scale benchmark dataset for Vietnamese-English machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4495–4503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving LLM abilities in idiomatic translation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 175–181, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Satyam Dwivedi, Sanjukta Ghosh, and Shivam Dwivedi. 2024. Navigating linguistic diversity: In-context learning and prompt engineering for subjectivity analysis in low-resource languages. *SN Computer Science*, 5(4):418.

Norbert Francis. 2023. Annals of vietnam: The preservation of a literary heritage. *Journal of Language, Literature and Culture*, 70(2):83–98.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation.

Dang Nguyen Giang. 2023a. Comparative images in vietnamese perception through idioms with comparisons. *Theory and Practice in Language Studies*, 13(9):2179–2185.

Dang Nguyen Giang. 2023b. Vietnamese concepts of love through idioms: A conceptual metaphor approach. *Theory and Practice in Language Studies*, 13(4):855–866.

Aaron Grattafiori et al. 2024. The llama 3 herd of models.

Ping Guo et al. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697.

Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023. Coverage-based example selection for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13924–13950, Singapore. Association for Computational Linguistics.

Nguyen Thi Bich Hanh, Dang Nguyen Giang, Ho Ngoc Trung, et al. 2023. Superlative degrees in vietnamese perceptions of humans through idioms with comparisons. *Eurasian Journal of Applied Linguistics*, 9(3):285–299.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. Towards effective disambiguation for machine translation with large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 482–495, Singapore. Association for Computational Linguistics.

Neil L Jamieson. 2023. *Understanding Vietnam*. Univ of California Press.

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

Thành Lã. 1995. Dictionary of current english-vietnamese idioms= từ điển thành ngữ anh việt thông dụng với 25000 thuật ngữ.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18554–18563.

Haoyu Liu et al. 2024. se2: Sequential example selection for in-context learning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5262–5284.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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

Raphaël Merx et al. 2024. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

Tuan Nguyen Minh, Phayung Meesad, and Huy Cuong Nguyen Ha. 2021. English-vietnamese machine translation using deep learning. In *International Conference on Computing and Information Technology*, pages 99–107. Springer.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Chinh Ngo, Trieu H. Trinh, Long Phan, Hieu Tran, Tai Dang, Hieu Nguyen, Minh Nguyen, and Minh-Thang Luong. 2022. Mtet: Multi-domain translation for english and vietnamese.

Dinh Hung Nguyen. 2014. *A Collection of Common Vietnamese-English Idioms, Proverbs, and Folk Verses*.

OpenAI et al. 2024. Gpt-4o system card.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Nghia Luan Pham, Thang Viet Pham, et al. 2023. A data augmentation method for english-vietnamese neural machine translation. *IEEE Access*, 11:28034–28044.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

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

Thi Minh Tran. 2024. Vietnamese heritage language: From silence to voice. In *Vietnamese Language, Education and Change In and Outside Vietnam*, pages 129–157. Springer Nature Singapore Singapore.

Elsa Vula and Nazli TyfekÃ. 2024. Navigating nonliteral language: The complexities of translating idioms across cultural boundaries. *Academic Journal of Interdisciplinary Studies*, 13.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

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

Thomas Wolf et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

An Yang et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pages 39818–39833. PMLR.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with qlora. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481.

A Comprehensive Results on LLMs

Model	#params	Methods	En→Vi			Vi→En		
			All	✓idioms	✗idioms	All	✓idioms	✗idioms
Qwen2.5	494M	✗	7.19	6.03	7.58	11.69	9.20	12.60
		✓	7.26	7.07	7.33	19.80	15.93	21.01
LLaMA-3.2	1.21B	✗	9.84	6.38	10.97	1.17	0.75	1.31
		✓	1.80	3.32	1.22	14.87	9.54	16.85
Qwen2.5	1.54B	✗	18.17	13.62	19.72	18.50	15.30	19.68
		✓	18.97	17.11	19.62	23.51	19.53	24.95
Gemma2	2.61B	✗	21.85	18.57	22.99	20.81	18.24	21.77
		✓	22.02	20.65	22.50	27.46	24.55	28.54
Qwen2.5	3.09B	✗	20.23	15.17	21.96	22.16	18.05	23.68
		✓	20.90	18.56	21.72	28.90	26.12	29.95
LLaMA-3.2	3.21B	✗	21.92	17.37	23.46	20.83	17.22	22.16
		✓	22.07	19.09	23.11	22.24	19.20	23.47
Qwen2.5	7.62B	✗	24.18	19.55	25.77	25.44	21.41	26.94
		✓	24.37	22.30	25.10	31.16	29.35	31.84
LLaMA-3.1	8.03B	✗	25.42	19.25	27.50	17.26	15.90	17.74
		✓	26.20	23.02	27.30	28.64	27.27	29.16
Gemma2	9.24B	✗	29.18	23.04	31.14	28.04	24.37	29.40
		✓	29.85	26.38	30.84	32.04	29.82	32.87

Table 7: BLEU score evaluation results of various open-resource LLMs, with (✓) and without (✗) the IDIAT framework, on the IDIAT benchmark dataset.

Model	#params	Methods	En→Vi			Vi→En		
			All	✓idioms	✗idioms	All	✓idioms	✗idioms
Qwen2.5	494M	✗	-59.84	-75.93	-51.69	0.46	-14.49	8.02
		✓	-62.49	-68.24	-59.58	30.83	14.44	39.13
LLaMA-3.2	1.21B	✗	-61.07	-74.85	-54.09	-93.28	-96.82	-91.48
		✓	-131.34	-122.46	-135.84	15.08	-18.92	32.29
Qwen2.5	1.54B	✗	-5.94	-18.23	0.28	29.46	15.34	36.60
		✓	-0.83	-9.86	3.74	48.39	34.69	55.32
Gemma2	2.61B	✗	19.02	5.02	26.10	36.60	21.04	44.47
		✓	22.68	15.14	26.50	51.82	35.48	60.09
Qwen2.5	3.09B	✗	4.73	-10.28	12.33	38.86	24.18	46.29
		✓	5.85	-3.10	10.38	52.61	36.42	60.80
LLaMA-3.2	3.21B	✗	15.54	0.98	22.91	33.08	18.09	40.67
		✓	17.90	9.17	22.31	48.45	35.47	55.02
Qwen2.5	7.62B	✗	14.31	2.24	20.42	45.29	31.93	52.05
		✓	15.18	8.56	18.53	55.34	46.08	60.02
LLaMA-3.1	8.03B	✗	31.81	17.76	38.92	23.66	14.91	28.08
		✓	35.27	24.23	40.86	55.22	43.44	61.18
Gemma2	9.24B	✗	45.02	33.38	50.90	48.55	34.76	55.53
		✓	48.10	41.18	51.60	58.24	46.69	64.08

Table 8: COMET score evaluation results of various open-resource LLMs, with (✓) and without (✗) the IDIAT framework, on the IDIAT benchmark dataset.

Besides the results on the commercial model, such as GPT-4o-mini, shown in the main Sections, we also present comprehensive evaluation results of various open-source LLMs on the IDIAT benchmark dataset. We compare the performance of different model sizes ranging from 0.5B to 9B parameters across three model families: Qwen2.5 (Yang et al., 2024), LLaMA-3.1 (Grattafiori et al., 2024), LLaMA-3.2 (Grattafiori et al., 2024), and Gemma2 (Team et al., 2024). Each model is evaluated with and without the IDIAT prompting framework, explicitly examining their performance on the idiomatic translation task.

As shown in Table 7, the integration of the IDIAT framework consistently improves translation quality across all model sizes and architectures. Looking at the overall BLEU scores, we observe several key trends. First, larger models generally perform better, with Gemma2-9B achieving the highest scores (29.85 for En→Vi and 32.04 for Vi→En with IDIAT). Second, the improvement from IDIAT is particularly pronounced for idiomatic expressions. Notably, the performance gap between idiomatic and non-idiomatic translations narrow significantly when IDIAT is applied, suggesting better handling of linguistic nuances.

COMET scores, illustrated in Table 8, show more dramatic improvements with IDIAT, particularly for Vi→En translation. The Gemma2-9B model demonstrates the most robust performance across all conditions, achieving positive scores even for idiomatic expressions. This suggests that larger models combined with IDIAT are particularly effective at handling the complexities of idiomatic language translation.

B Results on Multilingual Idiomatic Translation

To further assess the effectiveness of the IDIAT framework, we conduct experiments on multilingual idiomatic translation using GPT-4o-mini. We compile a multilingual evaluation set by collecting 10 idiomatic samples for each language pair, resulting in a total of 50 samples. The selected languages cover a broad spectrum of resource availability, ranging from extremely low-resource languages like Slovenian and Finnish, to low-resource languages like Thai, and mid-resource languages like Korean and Japanese.

Languages	**N.o. Speakers Worldwide**	**Methods**	**Source→En**	**En→Source**
Japanese	128M+	✗	24.63	20.57
		✓	$24.74_{\uparrow 0.11}$	$25.50_{\uparrow 4.93}$
Korean	77M+	✗	36.87	27.04
		✓	$42.02_{\uparrow 5.15}$	$30.47_{\uparrow 3.43}$
Thai	60M+	✗	11.30	42.50
		✓	$32.34_{\uparrow 21.04}$	$67.94_{\uparrow 25.44}$
Finnish	5.5M+	✗	37.53	32.89
		✓	$79.68_{\uparrow 42.15}$	$62.36_{\uparrow 29.47}$
Slovenian	2.5M+	✗	20.26	25.69
		✓	$29.13_{\uparrow 8.87}$	$49.01_{\uparrow 23.32}$

Table 9: Multilingual test results on X↔English, which X includes Japanese, Korean, Thai, Finnish, and Slovenian on BLEU score. Note that character-based language (Japanese, Thai, Korean) samples are assessed on character-based BLEU.

Table 9 presents BLEU scores for multilingual idiomatic translation between English and five languages: Japanese, Korean, Thai, Finnish, and Slovenian. Across all languages, the improved method consistently outperforms the baseline. These results highlight the effectiveness of the enhanced approach in handling idiomatic expressions across diverse linguistic structures, with especially strong performance in languages with smaller speaker populations, such as Finnish and Slovenian.

C Idiom Complexity Analysis

We extend the result analysis on the idioms' complexities, based on three aspects that can be taken into account, such as "Semantic Opacity", "Common Usage", and "Cultural and Linguistic Equivalence". The color-coded texts indicate the quality of translation:

Text: Accurate translation with equivalent idioms.
Text: Generally correct translation with similar meanings.
Text: Incorrect translation or literal translation (resulting in different meanings).

Note that the "ID" value in these tables represents the sample ID from the IDiAT benchmark dataset, which is included as part of this paper.

C.1 Semantic Opacity

Opaque Idioms have meanings unrelated to their individual words. **Semi-Opaque Idioms** show some link between the parts and the whole, while **Transparent Idioms** have meanings that can be reasonably inferred from the words themselves.

	Type: Opaque Idiom
ID: 716	En: **His heart was in boots** as he waited for news of the accident. Vi: Anh ấy **bồn chồn ruột gan** khi đợi chờ tin tức về vụ tai nạn.
en-vi	Zero-shot: **Trái tim anh nặng trĩu** khi chờ tin về vụ tai nạn. IDiAT: Trái tim của anh **bồn chồn ruột gan** khi chờ tin về vụ tai nạn.
vi-en	Zero-shot : He was **anxious and restless** while waiting for news about the accident. IDiAT: He was **restless with worry** as he awaited news about the accident.

	Type: Semi-opaque Idiom
ID: 660	En: For many years she was a **voice in the wilderness** protesting against child labour. Vi: Nhiều năm nay bà ấy đã **đơn thương độc mã** trong việc phản đối sử dụng lao động trẻ em.
en-vi	Zero-shot: Trong nhiều năm, cô là một **tiếng nói trong hoang dã** phản đối lao động trẻ em. IDiAT: Trong nhiều năm, cô là **một tiếng nói đơn độc** phản đối lao động trẻ em.
vi-en	Zero-shot: For many years, she has been **fighting alone** against the use of child labor. IDiAT: For many years, she has **been all alone** in her opposition to the use of child labor.

	Type: Transparent Idiom
ID: 959	En: The students decided to join together in order to present their grievances to the faculty, since **union is strength**. Vi: Các sinh viên quyết định đoàn kết lại để đưa những bức xúc của họ lên khoa, vì **đoàn kết là sức mạnh**.
en-vi	Zero-shot: Các sinh viên quyết định hợp tác với nhau để trình bày những phàn nàn của họ với giảng viên, vì **đoàn kết là sức mạnh**. IDiAT: Các sinh viên quyết định hợp tác với nhau để trình bày những phàn nàn của họ với giảng viên, vì **đoàn kết là sức mạnh**.
vi-en	Zero-shot: The students decided to unite to bring their grievances to the department, because **unity is strength**. IDiAT: The students decided to unite to bring their grievances to the faculty, as **unity is strength**.

For opaque idioms, zero-shot translations using GPT-4o-mini fail to produce accurate translations in the En→Vi direction, whereas IDiAT succeeds by employing equivalent idioms in En→Vi and conveying equivalent meanings in Vi→En. In the case of semi-opaque idioms, both methods face challenges when translating from Vi→En; however, they manage to generate relatively accurate translations in the En→Vi direction. For transparent idioms, both methods perform well, providing correct translations and using appropriate equivalent idioms in both translation directions.

C.2 Common Usage

Common Idioms are more likely to be memorized by the model, whereas **Rare Idioms** may be mistranslated or omitted.

Type: Common Idioms

ID: 250

En: There were smiles all round when the contract was signed - it was a **win-win** situation.

Vi: Mọi người đều rất vui vẻ khi hợp đồng được kí kết - đó là hợp đồng cho **đôi bên cùng có lợi** mà.

en-vi

Zero-shot: Có nụ cười rạng rỡ khắp nơi khi hợp đồng được ký - đó là một tình huống **đôi bên cùng có lợi**.

IDiAT: Khi trận đấu còn một giờ nữa kết thúc, huấn luyện viên đã quyết định thay đổi hậu vệ. **Thay ngựa giữa dòng**.

vi-en

Zero-shot: Everyone was very happy when the contract was signed - it was a **mutually beneficial** agreement.

IDiAT: Everyone was very happy when the contract was signed - it was a **win-win** situation.

Type: Rare Idioms

ID: 657

En: **When Hai fell in love with Ha, he used to tell Ha that she's the apple of his eyes**. But as they broke up, he insulted Ha with bad words. He's a bad guy.

Vi: Khi còn yêu nhau, Hải thường nói với Hà rằng Hà là **tất cả tình yêu của anh ta**. Nhưng khi họ chia tay nhau, anh ta đã lăng mạ Hà. Anh ta đúng là kẻ tồi tệ.

en-vi

Zero-shot: Khi Hải yêu Hà, anh thường nói với Hà rằng cô là **ánh sáng trong mắt anh**. Nhưng khi họ chia tay, anh đã xúc phạm Hà bằng những lời lẽ xấu. Anh ấy là một người xấu.

IDiAT: Khi Hải yêu Hà, anh thường nói với Hà rằng cô là **thứ yêu quý nhất của anh**. Nhưng khi họ chia tay, anh đã xúc phạm Hà bằng những lời lẽ xấu. Anh ấy là một người xấu.

vi-en

Zero-shot: When they were in love, Hải often told Hà that she was **his everything**. But when they broke up, he insulted Hà. He is truly a terrible person.

IDiAT: When they were in love, Hai often told Ha that she was **his everything**. But when they broke up, he insulted her. He is truly a terrible person.

For common idioms, the zero-shot method is capable of using equivalent idioms in the En→Vi direction but tends to only convey the general meaning in Vi→En, often omitting the use of equivalent English idioms even when they exist. In contrast, IDiAT performs well in both directions, preserving idiomatic expressions accurately. When it comes to rare idioms, the zero-shot method fails to produce accurate translations in En→Vi and only provides a relatively correct rendering in Vi→En. IDiAT, however, delivers relatively accurate translations in both directions. This discrepancy is exemplified by the idiom "apple of someone's eye," classified as a rare idiom[10].

[10] According to IDP: https://ieltskorea.org/korea/prepare/article-idioms-a-z-explained-5

C.3 Cultural and Linguistic Equivalence

Type: Direct Equivalent

ID: 290

En: When the match was one hour end, the coach decided to chang the defender. **Changing horses in midstream**.

Vi: Khi trận đấu đã diễn ra được một giờ, huấn luyện viên quyết định thay hậu vệ. Đó là **thay ngựa giữa dòng**.

en-vi

Zero-shot: Huấn luyện viên quyết định thay đổi hậu vệ khi trận đấu chỉ còn một giờ nữa, một tình thế được ví như **thay ngựa giữa dòng**.

IDiAT: Khi trận đấu còn một giờ nữa kết thúc, huấn luyện viên đã quyết định thay đổi hậu vệ. **Thay ngựa giữa dòng**.

vi-en

Zero-shot : When the match had been going on for an hour, the coach decided to substitute the defender. It was like **changing horses in midstream**.

IDiAT: When the match had been underway for an hour, the coach decided to substitute the defender. It was like **changing horses in midstream**.

Type: Near Equivalent

ID: 527

En: When the teacher was so angry that she shouted, all the students were **as quite as a mouse**.

Vi: Khi cô giáo giận giữ tới mức phải quát lên, tất cả học sinh đều **ngồi im như thóc.**

en-vi

Zero-shot: Khi cô giáo tức giận đến mức la hét, tất cả học sinh đều **im lặng như chuột**.

IDiAT: Khi cô giáo tức giận đến mức la hét, tất cả học sinh đều **im như thóc**, không ho he.

vi-en

Zero-shot: When the teacher got so angry that she had to shout, all the students **sat still as rice**.

IDiAT: When the teacher got so angry that she had to shout, all the students **sat still as if they were frozen**.

Type: No Equivalent

ID: 529

En: Hai is working as a restaurant manager. He wants to be a hotel manager. He is goingto quit his job in the restaurant to apply to be a manager of a luxury hotel. His friends advise him that **a bird in the hand is worth two in the bush**.

Vi: Hải đang làm quản lí nhà hàng. Anh ấy muốn trở thành quản lí khách sạn. Vì vậy anh ấy quyết định bỏ công việc hiện tại và nộp đơn vào khách sạn sang trọng. Bạn bè khuyên anh ấy **không nên thả mồi bắt bóng**.

en-vi

Zero-shot: Hai đang làm quản lý nhà hàng. Anh ấy muốn trở thành quản lý khách sạn. Anh ấy sắp nghỉ việc ở nhà hàng để nộp đơn làm quản lý cho một khách sạn sang trọng. Bạn bè anh khuyên rằng **có một con chim trong tay còn hơn hai con chim trong bụi**.

IDiAT: Hai đang làm quản lý nhà hàng. Anh ấy muốn trở thành quản lý khách sạn. Anh ấy sẽ nghỉ việc ở nhà hàng để xin làm quản lý cho một khách sạn sang trọng. Bạn bè anh khuyên rằng "**không nên thả mồi bắt bóng**".

vi-en

Zero-shot: Hai is currently working as a restaurant manager. He wants to become a hotel manager. Therefore, he decided to quit his current job and apply to a luxury hotel. His friends advised him not to let go of **what he has for something uncertain**.

IDiAT: Hai is currently managing a restaurant. He wants to become a hotel manager. Therefore, he decided to quit his current job and apply to a luxury hotel. His friends advised him that **a bird in the hand is worth two in the bush**.

Direct Equivalents are idioms with the same form and meaning in both languages. **Near Equivalents** need slight changes to convey the meaning, while **No Equivalent** means the idiom must be paraphrased to preserve its sense.

Both methods provide accurate translations for direct equivalents. However, for near equivalents, the zero-shot method exhibits shortcomings: in the Vi→En direction, it produces a literal translation that fails to convey the intended meaning, while in the En→Vi direction, it translates the word "mouse" as "chuột," which, although correct, is not an idiomatic expression in Vietnamese. In contrast, IDiAT demonstrates superior performance by translating correctly in En→Vi and preserving the intended meaning relatively well in Vi→En. In cases with no direct equivalent, the zero-shot method fails in both directions, whereas IDiAT effectively manages the translations by employing appropriate idiomatic expressions.

D Unseen-idiom Translation Analysis

For idioms absent from the context database (Section 2.2), IDiAT uses the LLM's contextual understanding to generate meanings or equivalent idioms (where feasible) for the translation prompt. Here is the analysis on the translations generated by the baseline (zero-shot prompting) and IDiAT with samples that contain unseen idioms (idioms that not appear in our idiom collection).

	Sample 1
ID: 651	En: I know these people, they're going to **lay into me big time.**
	Vi: Anh hiểu họ mà, họ sẽ dập cho anh **tơi bời khói lửa.**
en-vi	Zero-shot: Tôi biết những người này, họ sẽ **chửi mắng tôi rất nặng nề**.
	IDiAT: Tôi biết những người này, họ sẽ **tấn công tôi một cách dữ dội**.
vi-en	Zero-shot: He understands them, they will **beat him to a pulp in the smoke and fire**.
	IDiAT: I understand them; they're going to **tear me apart** with their criticism.

	Sample 2
ID: 993	En: She alights on the petals, drinks the nectar, and **takes off unscathed**.
	Vi: Nó đáp xuống cánh hoa, hút mật, và bay đi **bình an vô sự**.
en-vi	Zero-shot: Cô hạ cánh trên cánh hoa, uống mật ngọt và bay đi mà **không bị tổn thương**.
	IDiAT: Cô ấy đậu trên cánh hoa, uống mật hoa, và bay đi mà **không bị thương**.
vi-en	Zero-shot: It landed on the petal, sucked nectar, and **flew away safely**.
	IDiAT: It lands on the petals, drinks the nectar, and flies away **safe and sound**.

For unseen idioms in the two evaluated samples, the zero-shot method fails to translate the first sample and provides only a relatively correct translation for the second, capturing the general meaning but omitting the use of the exact idioms. In contrast, IDiAT outperforms the zero-shot method by successfully incorporating equivalent idioms in the Vi→En translations, even when these idioms are not explicitly present in the idiom collection but do appear in the evaluation test set. This improved performance is attributed to the Fuzzy Matching component in the IDiAT framework (Section 3.2), which enables the retrieval of equivalent or closely related idiom descriptions from the idiom collection to support more accurate prompting.

E Settings

In our experiments, we set the temperature parameter to 0 for GPT-based models and 0.1, the minimum allowable value, for open-source LLMs to ensure deterministic and consistent outputs. The maximum sequence length is fixed at 2048 tokens. All GPU-intensive experiments are performed on a single NVIDIA A6000.

For GPT-4o-mini, we access the model via the OpenAI API[11], while open-source LLMs were utilized through the HuggingFace Transformers library (Wolf et al., 2020), using checkpoints publicly available on the HuggingFace[12] Model Hub.

F Overlap Between Benchmark Idioms and Curated Collection

To evaluate the alignment between our benchmark idioms and the curated idiom collection, we analyze the overlap based on unique idiom occurrences. The collection comprises 2,493 English idioms and 2,432 Vietnamese idioms, while the benchmark test set includes 322 English and 174 Vietnamese idioms. Below are the exact matches between the test set and the collection:

- English: 162 out of 322 (50.31%)
- Vietnamese: 139 out of 174 (79.89%)

These overlap figures are based on exact string matches. However, idioms frequently appear in multiple surface forms, such as "bite one's tongue" vs. "bite his tongue", which can obscure underlying semantic matches. This variability is particularly notable in English but is also present in Vietnamese, as documented by Dang (2011). Consequently, while exact-match statistics provide a conservative estimate, the actual semantic coverage of the collection is likely higher.

G Prompts

G.1 Relevant Exemplar Generation

To generate relevant exemplars, we use a specific prompt, which is designed to generate multiple related yet distinct sentences in the source language. These generated sentences are followed by their translations into the target language. The obtained data pairs must adhere strictly to the specified dictionary format.

Task: Given a sentence in {src_lang}, generate 5 related but different sentences in {src_lang}. Then, translate each sentence into {tgt_lang}.

Each generated pair should be a dictionary with two keys: '{src_lang}' and '{tgt_lang}'. Ensure the format is strictly as follows:

```
[
"{src_lang}": "generated {src_lang} text",
"{tgt_lang}": "translated {tgt_lang} text"
]
```

Input:
{src_lang}: {src_text}

Please strictly follow the specified format, ensuring the {src_lang} and {tgt_lang} texts are both closely related to the original input.

G.2 Idiom Description Generation

For the idiom description generation, we ask the LLM to translate idioms from the source language to their equivalent in the target language while preserving their meaning. A natural and contextually accurate translation is provided if no equivalent idiom exists.

Task: Translate the given idiom, which is used in the input, from {src_lang} to its equivalent idiom in {tgt_lang}, preserving its meaning. If no equivalent idiom exists, provide a natural translation in {tgt_lang} language that conveys the same meaning (not a literal translation).

Input: {src_text}

Idiom: {idiom_src_text}

[11] https://platform.openai.com/docs/api-reference

[12] https://huggingface.co/models

G.3 Topic Description Generation

In this prompt, the LLM is asked to identify the topics of a given sentence in the source language using concise keywords. The output provides a brief yet informative topic description for the input sentence.

Task: Given a sentence in {src_lang}, use a few words to describe the topics of the following input sentence.

Input: {src_text}

Topic(s): topic1, topic2,...

G.4 LLM-based Demonstration Generation

We leverage CoT-inspired prompting to guide LLMs in generating idiom-focused demonstrations that reflect contextual reasoning. These outputs serve as targeted demonstrations within our ICL setup.

Task: Translate the given idiom, which is used in the input, from {src_lang} to its equivalent idiom in {tgt_lang}, preserving its meaning. If no equivalent idiom exists, provide a natural translation in {tgt_lang} that conveys the same meaning (not a literal translation).

Input: {src_text}
Idiom: {src_found_idiom}

Instructions:
1. Only translate the idiom, not the whole sentence.
2. Do not include any additional explanations, comments, or other text outside the translation.

H Human Evaluation

H.1 Question Template

For the human evaluation section, each annotator is asked to choose the best among the three obtained from three different methods.

Task: Choose the best translation of the source text, given its contained idiom and reference translated text in the target language:

Source text: {src_text}

Idiom: {idiom_src_text}

Reference text: {tgt_text}

[1] Translation from the Topline
[2] Translation from the Baseline
[3] Translation from the IdiAT

Your choice is: {Choose one of the above}

H.2 Annotation Guidelines

To ensure the quality of this assessment, we give annotators the guidelines along with the evaluation criteria. Note that if multiple translations are identical or completely matched, all of them will be labeled as the best translation. Then, we calculate the average scores of all annotators, which are the results listed in Table 6.

STEP 1: Familiarize Yourself with the Context
Carefully read the following elements:
Source Text: The original text in the source language.
Source Idiom: The idiomatic expression in the source text.
Reference Translation: The translation of the source text in the target language, provided for reference. Analyze how the **Source Idiom** is translated in the Reference Translation to understand its expected meaning or equivalent expression.

STEP 2: Review the Provided Translations
Assess the quality of the three translations in [1], [2], and [3].

STEP 3: Choose the Best Translation
Select the translation that best conveys the meaning and essence of the **Source Idiom** in the target language. Record your choice in the **Answer** column as follows:

- If there is one clear best translation, write the corresponding number (e.g., 1).
- If two translations are equally the best, write both numbers separated by a comma (e.g., 1,2).

STEP 4: Priority Guidelines for Selecting the Best Translation
Idiomatic Accuracy: Prioritize translations that accurately convey the **Source Idiom** as an equivalent idiom in the target language.
Idiomatic Meaning: If no translation provides an equivalent idiom, choose the one that best conveys the idiom's meaning naturally. Use a dictionary to confirm the idiom's meaning if needed.
Overall Meaning: If none of the translations adequately translate the idiom or its meaning:

- Consider the **Source Text** and its overall message.
- Select the translation that best preserves the overall meaning.
- Disqualify translations that add irrelevant information or omit key details.