# Don't Take it Literally!
# Idiom-aware Translation via In-context Learning

**Anonymous ACL submission**

## Abstract

The translation of idiomatic expressions often results in misunderstandings and inaccuracies, affecting both everyday communication and machine translation. This paper introduces **Idi**om-**a**ware **T**ranslation (IDIAT), a novel framework designed to enhance idiomatic translation. As part of this work, we curate a high-quality Vietnamese-English idiom collection to provide contextual support for in-context learning (ICL) during translation. Additionally, we present the IDIAT evaluation benchmark, which includes both idiomatic and non-idiomatic text pairs to assess general translation quality and idiomatic translation performance. By leveraging ICL in large language models, IDIAT enhances few-shot demonstrations with idiom and topic descriptions, improving translation accuracy. Empirical results demonstrate that IDIAT outperforms traditional methods while requiring fewer data samples, and human evaluations confirm its effectiveness. This work 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 data and code used in this paper are publicly available for research purposes[1].

## 1 Introduction

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 (NMT), 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). This issue is illustrated in Figure 1, which contrasts the shortcomings of literal translation with the effectiveness of the idiomatic 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; 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 challenges of idiomatic translation, particularly in low-resource languages like Vietnamese, we propose a novel framework called

[1] https://anonymous.4open.science/r/IDiAT

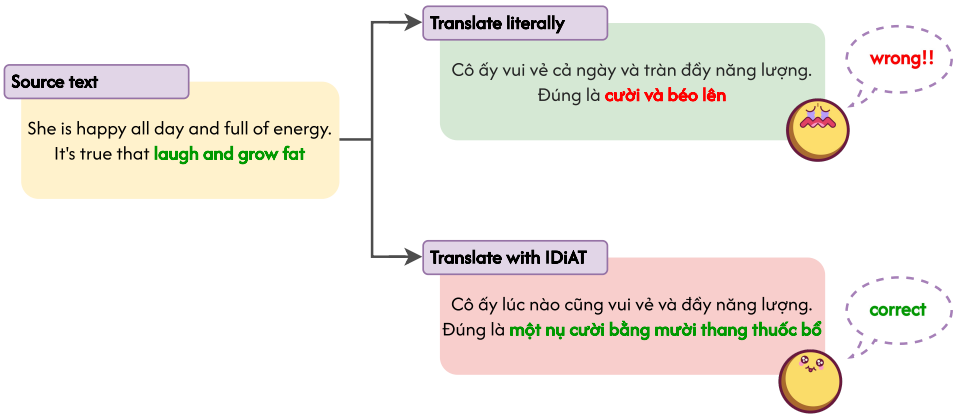


Figure 1: The Problem of Idiomatic Translation. While the literal translation of the idiom "laugh and grow fat" produces an incorrect and unnatural result in Vietnamese, the IDIAT framework captures the idiomatic meaning, yielding a culturally appropriate and accurate translation.

IDIAT. This harnesses the power of ICL in LLMs to convey the meanings of idioms in the target language accurately. IDIAT integrates three key components: few-shot demonstrations, idiom descriptions, and topic descriptions, which enhance translation performance, particularly for idiomatic expressions. By incorporating contextual information and relevant examples, IDIAT seeks to improve 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 can be summarized in three main key points:

- We introduce the IDIAT framework, which leverages the strengths of LLMs' in-context learning to improve idiomatic translation.
- We release the first evaluation benchmark specifically designed for idiom-aware translation in Vietnamese-English. Also, we present a high-quality idiom collection with equivalent pairs.
- We provide empirical evidence demonstrating the effectiveness of our approach in enhancing idiomatic translation through extensive experiments, showcasing significant improvements in translation quality across evaluation metrics.

## 2 Data Creation

### 2.1 IDiAT Benchmark Evaluation

Recognizing the scarcity of idiomatic expressions in existing Vietnamese-English translation evaluation datasets, we develop a high-quality evaluation set to assess general translation performance and idiomatic translation ability. To construct this set, we first filter the test set from the PhoMT dataset (Doan et al., 2021) to identify and select instances containing idioms. Additionally, we enrich the dataset by collecting more examples from official bilingual Vietnamese-English idiom reference textbooks. The resulting evaluation set contains 1,000 samples, with their distribution detailed in Table 1.

| Source | Have idiom | No idiom |
|---|---|---|
| PhoMT (Doan et al., 2021) | 181 | 664 |
| Textbooks[5] | 155 | 0 |
| **Total** | 336 | 664 |

Table 1: The distribution of 1,000 instances in the IDIAT benchmark evaluation test set taken from PhoMT dataset and some available Textbooks.

### 2.2 Idiom Collection

Previous research on enhancing idiomatic translation, such as IdiomKB (Ghazvininejad et al., 2023), figured out that prioritizing context awareness and using idiom descriptions in prompting provides a more comprehensive understanding of idiomatic expressions. Inspired by this, we propose constructing a comprehensive collection of Vietnamese idioms paired with their equivalent English translations to be used for idiomatic translation via ICL. The final dataset comprises 5,000 idiom pairs[6], each carefully validated to maintain equivalency between the source and target languages. This resource facilitates evaluation and contributes

[6]We crawled Vietnamese idioms and their equivalent English idioms from official bilingual Vi-En textbooks of idioms.

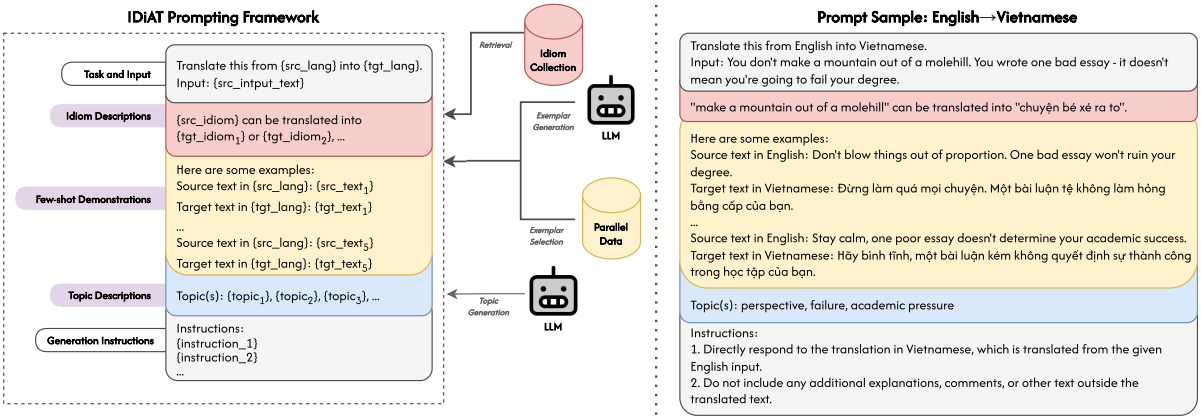


Figure 2: The IDIAT Prompting Framework consists of five key components: (1) *Task and Input*, which defines the task and input for the LLM; (2) *Few-shot Demonstrations*, providing exemplar translations to guide the model; (3) *Idiom Descriptions*, offering idiomatic translations for nuanced understanding;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

to advancing research on idiomatic translation for low-resource language pairs.

## 3 IDiAT: Idiom-aware Translation

In this study, we propose IDIAT, a framework designed 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 the entire framework, highlighting the flow of information between its key components.

### 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 (Gupta et al., 2023; Ye et al., 2023; Liu et al., 2024). This work explores multiple exemplar selection approaches, including Random Sampling, SBERT Similarity Ranking, and BM25 Ranking, to retrieve relevant examples from a large-scale existing dataset. Moreover, inspired by the chain-of-thought prompting technique (Wang et al., 2023; Wei et al., 2022b; Chu et al., 2024), which has proven effective in expanding the prompt context through LLMs themselves, we ask LLMs to generate relevant samples to assess their language understanding capabilities.

- **Random Sampling.** This method randomly selects a subset of translation examples from a larger dataset, which, while simple, can introduce variability in quality depending on the examples chosen.

- **SBERT Similarity Ranking.** This approach uses Sentence Transformers (SBERT) (Reimers et al., 2019) to compute semantic similarity scores between the input text and potential demonstration examples, enabling the model to rank and leverage the most relevant translation pairs to inform its output.

- **BM25 Ranking** (Robertson et al., 2009) is a probabilistic retrieval model that ranks documents based on their relevance to a query. In this context, it ranks translation examples based on their similarity to the input text, ensuring that the most contextually appropriate examples are presented to the LLM prompt.

- **LLM-generated Demonstrations.** This method involves generating demonstration examples using the LLM itself. By prompting the model to create its examples, we can obtain tailored translations that reflect its understanding of idiomatic expressions.

### 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 includes three different techniques for retrieval, including:

- **Exact Matching.** This method retrieves idioms matching the input idiom, ensuring precise equivalence.
- **Fuzzy Matching with Threshold.** This approach retrieves similar idioms, not identical, using a similarity threshold[7], making it suitable for cases with idiom variants.
- **BM25 Ranking.** Similar to its use in few-shot demonstrations, BM25 is employed here to rank idioms based on their relevance to the input idiom, facilitating the retrieval of contextually appropriate equivalents.

In addition, on the target language side, since an idiom may have multiple equivalent expressions, we employ two strategies to incorporate these target-language idioms into the prompt.

- **Use All.** This method retrieves all matching idioms from the collection and uses them in the translation prompt.
- **Use Top-1 by SBERT.** This approach uses a multilingual Sentence Transformer (Reimers and Gurevych, 2020) to compute cross-lingual similarity between the source and target idioms, selecting the top-ranked equivalent based on similarity scores.

For the idiom description generated by the LLM, 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.

[7]The threshold in this research is 0.7.

## 4 Experiments and Results

### 4.1 Settings

In this section, we outline the experimental settings used to evaluate the performance of our proposed framework, IDIAT, in the context of idiomatic translation.

**Model.** We use GPT-4o-mini, a compact version of GPT-4o (OpenAI et al., 2024), optimized for efficiency and strong NLP performance. The temperature is set at 0 for deterministic generations, and the sequence length is capped at 2048.

**Data.** The evaluation is conducted on the IDIAT benchmark dataset, described in Section 2, which includes idioms and non-idioms sourced from the PhoMT dataset and various textbooks.

An equivalent idiom collection is also constructed, containing 5,000 instances from specialized idiom textbooks.

For the few-shot demonstration retrieval in this study, we use a subset of 100K instances of the training set due to the large scale of the original dataset, which causes computational inefficiencies and extended processing times.

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

**Baseline.** We employ zero-shot prompting for the baseline. This approach allows us to assess the performance of our model without any prior fine-tuning on the specific idiomatic translation task nor in-context content for the prompting, providing a clear comparison against our proposed methods.

**IDIAT.** The proposed framework incorporates several key components to improve the translation performance of LLMs, particularly for idiomatic translation. It is noteworthy that processes requiring GPU computation are performed on a single NVIDIA A6000.

### 4.2 Evaluation Metrics

**Automated Metrics.** To assess the translation performance, we utilize two key metrics: sacreBLEU (Post, 2018)[8] and COMET (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,

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

| Methods | | En→Vi | | | | | | Vi→En | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | All | | ✓idioms | | ✗idioms | | All | | ✓idioms | | ✗idioms | |
| | | BLEU | COMET | BLEU | COMET | BLEU | COMET | BLEU | COMET | BLEU | COMET | BLEU | COMET |
| Topline: 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 | | | 31.27 | 54.99 | 30.48 | 46.72 | | |
| | 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 | N/A | | 31.16 | 54.80 | 30.07 | 46.15 | N/A | |
| | 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: Performance comparison on the IDIAT benchmark test set. Results are shown for all data ("All"), idiom-containing subsets ("✓ idioms"), and non-idiom subsets ("✗ idioms"). Bolded values indicate the best-performing method for each component tested across multiple approaches. Additionally, bolded 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 due to excluded idiom descriptions.

COMET provides a deeper assessment of semantic alignment, making it particularly effective for capturing the nuances of idiomatic expressions.

**LLM-based Metric.** Utilizing LLMs as evaluators for assessing the translation quality of idiom expressions 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[9].

**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 B, and asked to select the best translation among three approaches (Topline, Baseline, and IDIAT). The results of this evaluation are averaged across annotators to provide a robust measure of translation quality.

### 4.3 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 subsets 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 framework, compared to those of the baseline, in enhancing translation quality, particularly for idiomatic expressions.

**The addition of idiom descriptions benefits 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 topline model, indicating a substantial improvement. Similarly, in the Vi→En direction, the BLEU score for idioms rises to 32.29, surpassing the topline score of 25.50.

Moreover, the COMET scores also reflect sub-

[9]We re-implement Li et al. (2024)'s prompt for the GPT-score.

| 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.

stantial gains. In the En→Vi direction, the COMET score reaches 52.90 with IDIAT, compared to 32.12 (Topline), 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 topline score of 31.55.

Even the method of using LLM-generated idiom descriptions, which typically show variability in performance, 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 consistently outperform the topline 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 topline 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 human-like quality.

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

***w/o topic.*** The removal of topic descriptions causes slight performance declines in BLEU and COMET, 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 Discussions

We evaluate the proposed method using various LLMs (0.5B–9B parameters) across different model families, detailed in Appendix C. Additionally, we apply our pipeline to multiple X↔English pairs, including mid-resource (Korean, Japanese) and low/extremely low-resource languages (Thai, Finnish, Slovenian), with results in Appendix D. Findings confirm IDIAT's robustness, consistently surpassing the baseline. This section further analyzes results via GPT-score, human evaluation, and translation quality.

| 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.

| 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 5: 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.

### 5.1 GPT-score

In this section, we calculate the GPT-score on 100 samples randomly selected from the IDIAT benchmark dataset for this experiment. Note that those 100 samples all contain idioms.

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[10], 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 B.

Table 5 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%).

This strong performance highlights IDIAT's ability to align with human preferences, particularly for idiomatic expressions. Its consistency across both directions underscores its versatility in idiomatic translation.

Interestingly, the Baseline surpasses the Topline, suggesting that zero-shot prompting, despite lacking explicit fine-tuning, leverages LLMs' generalization abilities for idiomatic expressions better than supervised models trained on conventional parallel data. This indicates that traditional fine-tuning may struggle with idiomatic translation when training data lacks sufficient idiomatic coverage, whereas LLMs benefit from diverse linguistic patterns learned during pre-training.

### 5.3 Translations in Comparison

Table 6 compares idiomatic translations from three methods (Topline, Baseline, and IDIAT) for Vi↔En. In Vi→En, IDIAT correctly translates "**Không nên cầm đèn chạy trước ô tô**" as "**Don't put the cart before the horse**," while the others provide incorrect literal versions. Similarly, in En→Vi, it translates "**paddle his own canoe**" as "**tự lực cánh sinh**," capturing the idiomatic meaning. 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.

## 6 Related Work

Recent advancements in natural language processing, particularly with the emergence of LLMs and ICL techniques, have led to significant progress in translation and idiomatic expression handling, as reviewed in this section.

[10]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 xuống 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 6: 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.

### 6.1 LLMs and ICL in Translation

LLMs, such as the GPT series (Moslem et al., 2023; He et al., 2024; Pang et al., 2024), 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). This capability is particularly valuable for handling idiomatic expressions, which are traditionally challenging for translation models (Donthi et al., 2024; De Luca Fornaciari et al., 2024; Li et al., 2024; Phelps et al., 2024). 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 (Cahyawijaya et al., 2024; Dwivedi et al., 2024).

### 6.2 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[11], 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 translation of Vietnamese idiomatic expressions.

## 7 Conclusions

This work has explored the potential of in-context learning to enhance idiomatic translation, demonstrating its effectiveness in disambiguation and contextual understanding. Our proposed framework, IDIAT, integrates idiom descriptions and topic descriptions in the context and collectively improves the LLMs to generate semantically and culturally relevant translations.

Beyond improving translation accuracy, this research leverages the strengths of LLMs and ICL to create a robust framework for addressing idiomatic complexities, paving the way for future research. Testing the IDIAT framework on other low-resource and highly low-resource languages could expand its applicability, contributing to more inclusive and effective translation systems that bridge linguistic and cultural gaps.

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

## 8 Limitations

This study has several 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 framework across broader contexts and languages.

## References


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.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 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.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut to crack: Idiom detection with conversational large language models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

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 Patel, Joon Doh, and Eid Rodan. 2024. Improving llm abilities in idiomatic translation. *arXiv preprint arXiv:2407.03518*.

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, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and He-Yan Huang. 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.

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, Jianfeng Liu, Shaohan Huang, Yuefeng Zhan, Hao Sun, Weiwei Deng, Furu Wei, and Qi Zhang. 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, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 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.

OpenAI, Josh Achiam, et al. 2024. Gpt-4 technical report.

Jianhui Pang, Fanghua Ye, Longyue Wang, Dian Yu, Derek F Wong, Shuming Shi, and Zhaopeng Tu. 2024. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *arXiv preprint arXiv:2401.08350*.

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

Dylan Phelps, Thomas M. R. Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.

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.

Nils Reimers et al. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

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.

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, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 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 Prompts

## A.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.

## A.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}

## A.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,...

# B Human Evaluation

## B.1 Question Template

For the human evaluation section, each annotator is asked to choose the best among the three ones 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}

## B.2 Annotation Guidelines

To ensure the quality of this assessment, we give annotators the guidelines along with 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 5.

**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.

## C 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.

## D 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.