Large Language Models for Persian-English Idiom Translation

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown superior capabilities in translating figurative language compared to neural machine translation (NMT) systems. However, the impact of different prompting methods and LLM-NMT combinations on idiom translation remains underexplored. This paper introduces two parallel datasets for Persian→English and English→Persian translations, with Persian idioms sampled from our PersianIdioms resource, a collection of 2,200 idioms and their meanings. Using these datasets, we evaluate various open- and closed-source LLMs, Google Translate as an NMT system, and their combinations. Translation quality is assessed through idiom translation accuracy and fluency. We also find that automatic evaluation methods like LLM-as-a-judge and BLEU are effective for comparing model performance. Our experiments reveal that Claude-3.5-Sonnet delivers outstanding results in both translation directions. For English→Persian, combining weaker LLMs with Google Translate improves results, while Persian→English translations benefit from single prompts for simpler models and complex prompts for advanced ones.[1]

1 Introduction

An idiom is a phrase or expression with a figurative meaning distinct from its literal interpretation. Idioms are commonly used in everyday language to convey ideas more vividly and often originate from cultural, historical, or social contexts, making them specific to particular languages or regions. Idiomatic expressions, including idioms and sayings, present significant challenges for NLP, particularly in translating between culturally distinct languages such as Persian and English.

Despite their prevalence in spoken language, state-of-the-art machine translation (MT) models struggle with translating idioms, often rendering them literally as compositional expressions (Raunak et al., 2023; Dankers et al., 2022). Early MT efforts attempted to address this problem using idiom dictionaries or direct substitution (Salton et al., 2014; Nagao, 1984). However, idioms evolve, vary by context, and even identical meanings can require different translations. E.g., the idiom "Keep at bay" has a different contextual meaning and therefore translation in these two sentences: (i) "The infection is kept at bay." meaning: "The infection is under control." (ii) "The fire keeps the wolves at bay." meaning: "The fire keeps the wolves away."

Recent large language models (LLMs) have enabled improvements in idiom translation compared to neural MT (NMT) systems (Raunak et al., 2023), due to their higher tendency towards non-literalness (Raunak et al., 2023) and greater paraphrastic capability (Hendy et al., 2023). However, no research has yet compared the performance of LLMs across different prompts, prompting techniques, and in combination with NMT models.

In this research, focusing on English-Persian translation, we try to fill several gaps. Since Persian datasets are limited in this context, we first introduce a comprehensive resource for idiomatic expressions in Persian (PersianIdioms). This resource captures idiomatic expressions and their meanings, including contextual usage examples. Additionally, we produce two parallel En→Fa and Fa→En datasets, each consisting of 200 sentences containing English and Persian idiomatic expressions. We then use these datasets to evaluate the performance of open-source—Qwen-2.5-72B (Team, 2024) and Command R+ (Cohere For AI, 2024)—, and closed-source—GPT-3.5 (OpenAI, 2023), GPT-4o-mini (OpenAI et al., 2024) and Claude-3.5-Sonnet (Anthropic, 2024)— LLMs, and Google Translate as an NMT system, and a combination of them in idiom translation across various settings.

We assess translation quality using two metrics,

[1] We will release all our resources and annotations upon publication.

idiom translation and fluency. Then, we explore suitable replacements for manual evaluation by calculating the correlation between scores from existing automatic evaluation approaches and manually obtained scores.

In summary, our main contributions are as follows. (i) A new resource for Persian idioms, PersianIdioms, which includes about 2200 idioms, with their meaning. A subset of 700 also contains example usage, which currently does not exist for Persian. (ii) New parallel Fa→En and En→Fa datasets (200 examples each) with inputs containing at least one idiom. (Persian sentences are from PersianIdioms, and English ones are from various sources including EPIE (Saxena and Paul, 2020), MAGPIE (Xu et al., 2024) and Abadis online dictionary. (iii) Conducting a comprehensive comparison of human evaluations versus LLMs-as-a-judge in assessing idiom translation quality, through exhaustive manual and model annotations of various translation outputs. (iv) Evaluating different prompting methods and a combination of LLMs and Google Translate on idiom translation and highlighting how they perform differently for Fa→and English En→Fa.

2 Related work

We first review some datasets for idiom expressions, then reference a few recent works focusing on LLMs for translation.

2.1 Idiom datasets

Xu et al. (2024) created the largest English idiom corpus to date, containing over 50K instances, by using a fixed idiom list, automatic pre-extraction, and a controlled crowdsourced annotation process, enabling further analysis of idiom usage across genres. Saxena and Paul (2020) compile the EPIE dataset of sentences containing highly occurring English idioms and idioms using StringNet. Kabra et al. (2023) create the MABL dataset covering the figurative language from 7 typologically diverse languages. The crowdsourced metaphors and idioms highlight cultural and linguistic variations. Liu et al. (2023) investigate the ability of multilingual language models (mLLMs) to reason with cultural common ground by using idioms and sayings as a proxy. They construct a new dataset called MAPS (MulticultrAl idioms and Sayings) covering 6 languages with idioms, conversational usages, interpretations, and figurative labels. Li et al. (2024) present a methodology for constructing a large-scale, multilingual idiom knowledge base by distilling figurative meanings from language models. Liu et al. (2022) introduce Fig-QA, a new task to test language models' ability to interpret figurative language. They crowdsource a dataset of over 10k paired metaphorical phrases with opposite meanings and literal interpretations. These works demonstrate techniques for compiling figurative language data across multiple languages. However, they are focused on English or non-Persian languages. A large-scale, Persian-specific idiom dataset is still lacking. This research applies similar techniques of utilizing existing resources and language model generation to create idiom data specifically for Persian.

2.2 Translation and LLMs

Jiao et al. (2023) demonstrate that ChatGPT competes well with commercial translation services like Google Translate for high-resource European languages but struggles with low-resource or distant languages. With the introduction of the GPT-4 engine, ChatGPT's translation capabilities have improved significantly, reaching a level comparable to commercial products, even for distant languages.

Moslem et al. (2023) evaluate GPT-3.5's performance across various translation tasks, including adaptive MT, comparing it with strong encoder-decoder MT systems. GPT-3.5 shows excellent results for high-resource languages but struggles with low-resource languages and certain tokenization issues with Arabic. Hendy et al. (2023) suggest that the increased tendency for paraphrasing in GPT translations may be beneficial for enhancing NMT models in the translation of figurative language. We validate this hypothesis empirically in our paper in the case of English and Persian translations. Yamada (2024) offer two prompts aimed at enhancing the quality of translations generated by ChatGPT. We will assess and contrast these prompts with our approaches. Raunak et al. (2023) propose novel evaluation metrics for measuring translation literalness and compare the performance of LLMs from the GPT series and NMT models in idiom translation, finding that translations produced by GPT models are generally less literal. Several other studies have also been dedicated to comparing the performance of a variety of different LLMs and NMTs for different languages (Castaldo et al., 2024; Zhu et al., 2024; Tang et al., 2024). Others have aimed to enhance LLM idiom translation through fine-

tuning (Stap et al., 2024) and augmenting them with knowledge bases (Li et al., 2024),

However, these efforts primarily focus on individual model performances, overlooking the effects of more complex prompting techniques and the potential advantages of combining these models. Additionally, the evaluation of idiom translation between English and Persian has remained unexplored. This work seeks to fill these gaps.

3 Datasets

In this section, we explain the process of building our datasets. Two parallel datasets are created for En→Fa translation and inverse. Besides, we collect a comprehensive resource for Persian idioms which is explained in the following section.

3.1 PersianIdioms

Here, we explain how we build PersianIdioms. Our data collection begins with extracting Persian idioms and their meanings from an online dictionary called Abadis[2]. For each idiom, we also gathered usage examples to provide contextual clarity, sourced from user-generated examples in Abadis. We changed these examples manually to have just one idiom and they are simplified to make the translation task more accurate. These examples are crucial for future testing of language models, allowing them to learn from actual idiomatic expressions in use. Notably, this comprehensive dataset of Persian idioms, their meanings, and contextual usage examples has never existed before, making it a valuable resource for the development and evaluation of language models for Persian.

Data verification Once the idioms are collected, they are annotated with their meanings and supplemented by examples. The annotated dataset then undergoes a thorough cleaning process to ensure both accuracy and consistency. To further enhance its quality, native Persian speakers review the idioms, verifying the accuracy of their meanings and the appropriateness of the contextual examples. This manual verification step is crucial for maintaining both linguistic integrity and cultural relevance.

As a result, the dataset comprises 2,200 idiom-meaning pairs, with 700 idioms further enriched with contextual examples. This dataset not only highlights the richness and complexity of Persian idiomatic expressions but also serves as a valuable resource for advancing NLP research in interpreting culturally nuanced language.

[2]https://abadis.ir/ (The Abadis website mentions that using the entries of this dictionary is permitted, if the resource is cited.)

Idiom	اب دوغ خیاری
Meaning	پیش پاافتاده / مبتذل
Meaning in English	low quality/tasteless
Example	هروقت می رم خونه شون همه پای تلویزیون نشسته ان و دارن یکی از این فیلم های اب دوغ خیاری رو تماشا می کنن.
Gold translation	Every time I go to their house, everyone is sitting in front of the TV watching one of those low-quality movies.

Table 1: Persian idiom details in dataset

3.2 Translation datasets

Fa→En From the idioms containing contextual examples in the Persian idiom dataset, we randomly select two hundred samples. Subsequently, a proficient translator produces English interpretations of these selected sentences, which are then reviewed and validated by another expert. You can see an example of our data in Table 1.

En→Fa In our initial data collection phase, we attempted to identify sentences containing idiomatic expressions from existing En→Fa parallel resources. However, we found that the Persian translations in these datasets were either automatically generated, derived from translations of English literature into Persian (Kashefi, 2020), or sourced from Wikipedia (Karimi et al., 2019). Each of these approaches posed significant challenges to our research objectives. Automatic translation by NMT models, often produces inaccurate results, especially for figurative language, which is the focus of this study. Translations derived from English literature tend to incorporate contextual references, such as character names, or modify sentence structures to enhance fluency in the target language. Wikipedia-sourced sentences largely lack complex, culturally specific idioms and primarily feature easily translatable expressions like “under pressure”.

Given these limitations, we opted for manual data collection. Drawing primarily from the EPIE and MAGPIE corpus, we carefully selected sentences that emphasize the challenges of translating idiomatic expressions, rather than overall sentence structure. A proficient translator then produced Persian renditions of these selected sentences, followed by a review and validation process conducted by

another qualified expert.

The result of these efforts is a dataset comprising 200 pairs of English sentences and their corresponding Persian translations. Since datasets containing English idioms and their meanings already exist, we avoided redundant work in this area.

4 Evaluation Methodology

In this section, we present two novel manual evaluation metrics and identify appropriate automatic metrics to serve as replacements using Spearman's Correlation and agreement percentage.

4.1 Manual evaluation

Metrics Using the MQM evaluation framework (Lommel et al., 2014), we devise two independent evaluation metrics: **fluency** and **idiom translation**. Idiom translation, a binary metric (0 or 1), assesses whether the translation preserves the idiom's meaning within the original sentence's context. Fluency, rated from 1 to 5, evaluates the syntactic and semantic correctness of the translation, assuming the idiom is correctly translated. Notably, idiom translation focuses solely on the semantic accuracy of the idiom, with grammatical errors affecting only the fluency score, not the idiom translation score.

We distill adequacy down to idiom translation for two reasons. First, idioms are the core of this study, and we aim to improve idiom translation without compromising overall performance. Thus, occasional non-idiom-related semantic errors only affect fluency scores. Second, our dataset consists of single sentences, that shift the translation challenge to the idiom itself. Therefore, the semantic accuracy of the entire sentence usually depends on the semantic accuracy of the idiom translation.

We chose binary labels over a 1-5 scale for idiom translation since idioms, being short phrases, rarely have partially correct translations. While a 1-5 scale might reflect how closely a translation aligns with the intended meaning, it is subjective and heavily influenced by factors like cross-linguistic transferability. The binary label simply checks whether the idiom's meaning is conveyed correctly. If the translation preserves the meaning but sounds unnatural, it is treated as a fluency issue, not an idiom translation error.

Annotation Three native Persian-speaking MSc students (some of them are the authors), fluent in English, were given detailed evaluation guidelines with examples and tasked with manually scoring the first 100 sentences from seven outputs generated by GPT-3.5, Google Translate, and their combination. The GPT-3.5 outputs were produced using the prompts outlined in Section 5.1. Idiom translation labels were decided by majority vote, and fluency scores were averaged across annotators' ratings. To ensure reliability, inter-annotator agreement was assessed for both idiom translation and fluency.

Fluency scores are highly skewed, with most labels between 3 and 5 and the 1-5 scale is subjective to an extent, resulting in low chance agreement, so we use Gwet's AC1 (Gwet, 2008) to assess inter-annotator agreement for fluency. Idiom translation labels are clearer and less subjective, allowing us to calculate agreement using Fleiss' Kappa. Additionally, we report observed fluency agreement based on a 1-point difference threshold and observed idiom translation agreement based on the proportion of sentences with matching idiom translation labels.

Table 2 shows inter-annotator agreement scores. For idiom translation in both directions, annotators align well, with Kappa exceeding 0.6 and high observed agreement. Fluency scores show moderate agreement based on Gwet's AC1, slightly higher for Persian. High observed accuracy further supports the reliability of these ratings. Given the task's subjectivity and complexity, these fluency agreement levels are acceptable.

Models	Agreement	
	En-Fa	Fa-En
Fluency (Gwen-AC1)	.45	.54
Fluency (Observed)	.84	.83
Idiom Translation (Fleiss Kappa)	.63	.68
Idiom Translation (Observed)	.74	.73

Table 2: Inter-annotator agreement for fluency and idiom-translation

4.2 Automatic evaluation

Our experiments test various prompting methods across different models. Manually scoring all configurations is laborious, so we seek suitable automatic metrics. To do this, we calculate Spearman's correlation between manual and automatic scores for the first 100 sentences from seven manually scored model outputs. The automatic metric with the highest correlation will be chosen as the best fit for assessing idiom translation and fluency.

Correlations are computed from overall scores derived from the first 100 sentences of each output. Thus, for each metric, we have a list of seven

	Metric	COMET	BERTScore	BLEU	GPT-4o
En→Fa	Fluency	.17	.89	**.96**	-.45
	IdiomT	.40	.18	-.03	**.85**
Fa→En	Fluency	.45	**.70**	.46	.15
	IdiomT	.35	.22	-.51	**.75**

Table 3: Spearman's Correlation between results obtained from automatic evaluation metrics and manually for En→ Persian and Fa→En. The best correlation for each row is in bold. IdiomT: idiom translation

		Human	GPT-4o
En→Fa	Human	.81	.76
Fa→En	Human	.73	.71

Table 4: Agreement % on idiom translation between human annotators and GPT-4o.

scores—one per output—on which correlations are based. For fluency, a model's score is the average of its 100 manual fluency ratings, while idiom translation accuracy is measured by the percentage of correctly translated idioms.

Existing automatic evaluation metrics To evaluate the quality of the translated text, we first use BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020), and COMET (Rei et al., 2020).

GPT-4o We follow the LLM-as-a-judge trend (Zheng et al., 2023; Li et al., 2024) and use GPT-4o to evaluate our outputs. We use the single answer and reference-guided grading as in Zheng et al. (2023). We experimented with GPT-4o and multiple prompts to maximize correlations. For En→Fa, the prompt "Is the idiom in this sentence correctly translated into English/Persian? Answer with just a number: 1 for yes and 0 for no. idiom: <idiom>, English sentence: <reference>', translation: <model translation>" yielded the best results. For Farsi, a lower resource language, including a gold translation as a reference, improves correlation, helping the model better assess the accuracy of idiom translations by providing additional guidance. However, for En→Fa, GPT-4o assigns a score of 1 only to translations closely resembling the gold standard, leading to false negatives due to the flexibility of idiom rewrites.

We also provide three examples to emphasize the importance of accurate idiom translation. For each model output, the overall idiom translation score is equal to the percentage of "correct idiom translations" predicted by GPT-4o. Finally, we set the temperature hyperparameter to 0.1 to minimize response variations.

4.3 Choosing suitable automatic metrics

Table 3 shows the correlation between manual and automatic evaluation metrics for idiom translation. BLEU penalizes non-literal translations, while COMET exhibits a stronger correlation with idiom translation performance, likely due to its consideration of both the source and gold translations. However, COMET's focus on idiomatic accuracy reduces its correlation with fluency, which is independent of idiom translation. As a result, Google Translate scores high on fluency despite translating idioms literally. Similarly, BLEU and BERTScore favor Google Translate, as it rarely paraphrases.

Interestingly, the correlation with fluency for Fa→En translations is lower for BERTScore and BLEU than for En→Fa. This suggests that Fa→En translations may involve more paraphrasing and structural changes. However, these correlations are not strong enough to draw definitive conclusions.

Ultimately, for En→Fa, the high correlation between BLEU and fluency, along with GPT-4o's evaluations for idiom translation, suggests that BLEU and GPT-4o are well-suited for ranking fluency and idiom translation performance, respectively. For Fa→En, BERTScore and GPT-4o are the most reliable metrics for ranking fluency and idiom translation performance.

We further examine GPT-4o's performance as a judge by calculating the agreement percentage between manually obtained idiom translation scores and GPT-4o labels. Table 4 compares the average agreement between human annotators and between GPT-4o and each annotator. The agreement between GPT-4o and human annotators approaches inter-annotator levels, suggesting GPT-4o performs comparably to humans and can serve as a reliable evaluation method for idiom translation.

5 Experiments

In this section, we outline the models and settings used in our translation experiments, followed by the results and their analysis.

5.1 Models and prompts

We use Google Translate (GT), and several open- and closed-source LLMs (GPT-3.5-turbo, QWEN-

2.5 with 72b parameters, Command R+ with 104b parameters, GPT-4o-mini, Claude-3.5-Sonnet), and a hybrid approach combining the mentioned LLMs with GT to generate translations.[3]. The hybrid approach (LLM + GT) first guides the LLM to identify and replace idioms with literal expressions, and then it applies GT to translate the resulting text into the target language.

The prompts used for LLMs for En→Fa translation are shown in Table 5. The second single prompt is taken from the prompts presented in Yamada (2024). Fa→En prompts replace "English" with "Persian" and vice versa, and "American" with "Iranian".

Accordingly, we consider five prompts grouped into three categories: (i) three single prompts (ii) one chain of thought (CoT) prompt, and (iii) one multiple prompt, which breaks down a single prompt into multiple independent steps. In the CoT setup, the three steps are provided as separate prompts, with each step and the model's response carried forward as chat history for the subsequent step. In contrast, the third category eliminates the reliance on chat history by using two independent prompts, where the answer to the first prompt is embedded within the second prompt itself. Initially, we employed a single prompt containing all three steps for CoT, but manual inspection revealed that breaking them into multiple prompts yields significantly better and more coherent results.

We set the temperature to 0.8 to balance reducing response variability while preserving some creative freedom. To prevent this variation from skewing evaluations, We ran our experiments with GPT models multiple times. Although individual sentence scores fluctuated between runs, the overall score remained consistent or changed only slightly, confirming that response variations did not significantly affect the final evaluation results.

5.2 Results: En→Fa

The left side of Table 6 presents the results of the automatic evaluation for En→Fa translation. As discussed in section 4.3 BLEU and GPT-4o scores are the most suitable metrics for ranking model performance in terms of fluency and idiom translation in this direction. Therefore, in this section, we will primarily focus on these two metrics.

Claude-3.5-Sonnet achieves the highest BLEU and GPT-4o scores, making it the most effective model overall. Our manual inspections reveal that this model not only excels in accurately identifying and understanding idioms but also effectively finds suitable Persian idiom replacements, contributing to its strong translation capabilities.

Google Translate consistently achieves higher BLEU scores than GPT-3.5, Qwen-2.5, and Command R+ across all prompts. When combined with Google Translate, these models show an increase in BLEU, benefiting from the strengths of both LLMs and NMT models. Notably, the GPT-4o score also improves significantly for Qwen-2.5 and GPT-3.5. For Command R+, the hybrid model's GPT-4o score nearly matches the model's best, while maintaining superior fluency. In contrast, GPT-4o mini and Claude-3.5 Sonnet outperform Google Translate in BLEU with certain prompts but see a performance decline when combined with it.

A key takeaway from these results is that weaker models like Qwen-2.5-72B, when combined with NMT models, can perform comparably to much stronger models such as Claude-3.5-Sonnet. This suggests that when an LLM's fluency is lower than that of an NMT model, combining the two can effectively close the performance gap with stronger LLMs. However, the efficiency and scalability of this approach warrant further investigation in future wor.

Across all models, the second and third prompts consistently outperform the first. The first prompt simply requests a sentence translation, increasing

Single prompts	Translate this sentence to Persian.
	Translate the following English text into Persian. Use natural expressions that can be understood by Persian speakers, unfamiliar with American Culture.
	Translate the following English text into Persian. Avoid word-for-word translations.
CoT Prompt	1) Identify the idioms in this sentence. 2) Replace the idioms with literal clauses. 3) Translate the literal sentence to Persian. Avoid word-for-word translation.
Multiple Prompts	Identify the idioms in this sentence and replace them with literal clauses. 2) Translate this literal sentence to Persian. Avoid word-for-word translation.

Table 5: The translation prompts used in our experiments.

[3]We utilized the OpenAI API and https://openrouter.ai/ to access these models, incurring an approximate cost of $60 in total for API usage.

	En→Fa				Fa→En			
	COMET	BERTScore	BLEU	GPT4o	COMET	BERTSCore	BLEU	GPT4o
GPT-3.5								
• Prompt1	82.6	82.3	11.9	63.5	**75.3**	**93.5**	23.2	36.0
• Prompt2	82.7	81.6	8.6	65.0	74.8	93.1	20.6	40.0
• Prompt3	**83.0**	82.3	11.1	65.5	74.8	93.1	21.8	**43.0**
• CoTPrompt	82.1	81.2	8.3	68.0	72.4	92.6	19.3	29.0
• MultiPrompt	81.4	80.7	7.9	72.0	71.6	92.6	18.4	30.0
• +GT	85.1	**84.6**	**19.3**	**79.0**	74.5	92.8	**25.3**	25.0
Qwen 2.5 72B								
• Prompt1	82.6	84.3	14.7	66.0	75.3	93.5	26.4	35.0
• Prompt2	83.0	82.1	13.4	72.0	76.5	93.4	23.8	**41.5**
• Prompt3	83.0	81.8	12.2	74.5	76.5	93.6	25.4	39.5
• CoTPrompt	80.0	79.5	7.2	72.5	75.8	93.6	**27.2**	34.5
• MultiPrompt	80.5	80.1	9.2	74.0	**76.7**	**93.7**	26.5	35.5
• +GT	**84.2**	**83.7**	**17.9**	**88.0**	74.0	93.0	24.6	24.0
GPT4o-mini								
• Prompt1	85.0	84.5	18.7	86.0	77.16	94.2	29.6	52.0
• Prompt2	85.5	84.5	**19.9**	90.0	79.4	94.4	27.0	56.0
• Prompt3	**85.8**	**84.6**	16.5	87.5	**79.6**	**94.3**	26.3	**62.0**
• CoTPrompt	84.5	83.2	15.1	86.0	79	93.9	25.0	52.5
• MultiPrompt	83.9	83.1	15.7	**91.0**	78.9	94.0	**30.0**	55.5
• +GT	84.7	84.0	17.3	81.5	79.0	93.7	26.9	54.0
Command R+								
• Prompt1	83.2	82.2	12.2	78.5	75.0	93.0	21.1	55.5
• Prompt2	82.6	81.5	10.5	75.0	76.7	93.2	21.2	57.0
• Prompt3	83.3	81.6	10.1	**83.5**	75.8	92.2	14.7	52.5
• CoTPrompt	78.7	78.8	6.9	69.5	70.9	91.3	13.8	**60.5**
• MultiPrompt	79.30	78.5	5.9	70.0	74.5	92.3	15.1	55.5
• +GT	**84.4**	**83.9**	**17.7**	81.5	**77.4**	**93.3**	**23.6**	57.0
Claude 3.5 Sonnet								
• Prompt1	85.1	84.6	**21.1**	91.0	79.7	94.6	**32.1**	68.0
• Prompt2	83.3	83.0	19.0	88.5	78.9	94.3	25.9	71.0
• Prompt3	**86.0**	**84.4**	20.8	93.5	79.2	94.3	24.7	70.5
• CoTPrompt	84.3	82.9	15.3	**94.0**	82.1	94.4	24.2	74.0
• MultiPrompt	83.5	82.9	17.0	90.0	**82.8**	**94.8**	29.8	**75.0**
• +GT	84.7	84.2	18.5	91.5	77.4	93.3	23.6	61.0
Google Translate (GT)								
• Direct Translation	81.1	83.7	17.6	52.0	73.9	93.0	26.1	21.0

Table 6: Results for automatic evaluation of different models and settings on 200 samples. The best overall scores are underlined, while the best scores across different configurations of a model are highlighted in bold.

the likelihood of literal idiom translations. In contrast, the second and third prompts provide more guidance, reducing literal translations but slightly compromising fluency. Our observations confirm that when prompts indicate certain elements can't be translated literally, models often paraphrase or over-explain idiom definitions, which can result in the translation straying from the original sentence, ultimately lowering BLEU scores.

For most models (except Command R+), CoT or multiple prompts, which break the task into smaller steps, improve GPT-4o scores but reduce fluency. It's important to note that n-gram-based metrics like BLEU are ill-suited for figurative language since they prioritize exact matches over semantic similarity, explaining low BLEU scores even when other metrics indicate better performance. Nevertheless, BLEU still serves as a useful metric for ranking fluency, as shown by its Spearman correlation with idiom translation accuracy.

Table 7 highlights the strengths and shortcomings of each translation method for GPT-3.5 through an example. Google Translate offers fluent translations but often renders idioms literally. Single prompts correctly detect idioms but produce unnatural definitions. Chain and multiple prompts translate metaphors more accurately but still struggle with fluency. Combining GPT-3.5 with Google Translate yields the most fluent translations while effectively capturing the meaning of idioms within their contextual sentences

Model	En →Fa Translation	Fa →En Translation
Sentence	Poor Mrs has lots of children and they were driving her up the wall!	من اعتماد کردم و حرف دلم را به او زدم اما او به هر کس رسید همه را روی دایره ریخت
Gold Translation	خانم بیچاره بچه های زیادی دارد که او را آزرده خاطر میکردند!	I trusted and spoke my heart to him, but he **revealed all my secrets to everyone he encountered.**
Google Translate	خانم بیچاره بچه های زیادی دارد که او را از دیوار بالا میبرند. Poor Mrs has lots of children that take her up the wall.	I trusted and spoke my heart to her, but she **threw everyone on the circle.**
Single Prompt	خانم فلانی بچه های زیادی دارد که دارند به اعصابش میزنند. Some Mrs has lots of children that are hitting her nerves.	I trusted him and opened up to him, but he **betrayed my trust by sharing everything with everyone.**
Chain Prompt	مادر بدبخت بسیاری فرزند داشت و آن‌ها باعث ایجاد تنش شدید برای او بودند. Poor mother had many children that had been causing her a lot of tension.	I trusted him and confided in him, but he **betrayed everyone and turned against all.**
Multi Prompt	خانم فقیر دارای بسیاری فرزند است و آن‌ها باعث ایجاد تنش شدید برایش می‌شوند! Poor (in Persian, the word used means impoverished, not unfortunate) Mrs has many children that have been causing her extreme tension!	I trusted him and spoke my heart out to him, but he **reached out to everyone and spread it on everyone's face.**
GPT-3.5 + Google Translate	خانم بیچاره بچه های زیادی دارد که او را اذیت می کردند! Poor mother had many children that are bothering her!	I trusted and told her my heart, but she **reached out to everyone and revealed everything.**

Table 7: Examples of En →Fa and Fa →En translations with back-translations to demonstrate translation quality.

5.3 Results: Fa→En

The right side of Table 6 shows the automatic evaluation results for Fa→En translation. As discussed in Section 4.3, BERTScore and GPT-4o are the most appropriate metrics for ranking model performance in fluency and idiom translation in this direction. Therefore, we focus primarily on these two metrics.

GPT4o-mini and Claude-3.5 Sonnet excel in idiomatic translation, outperforming other models across all metrics. GPT4o-mini with Prompt 3 and Claude-3.5 Sonnet with MultiPrompt deliver the most accurate, contextually aware, and fluent translations. Advanced techniques like multiple prompts and CoT further enhance idiomatic translation in larger models, such as Claude-3.5 Sonnet and Command R+, improving the disambiguation of idiomatic expressions and overall translation quality.

Our results also show that GPT-3.5, GPT-4o-mini, and Qwen-2.5 perform best with single prompts but struggle with more complex ones. These models tend to become confused in MultiPrompt and CoTPrompt setups, often failing to accurately identify or translate idioms. Manual evaluation of GPT-3.5 highlights two key challenges in Persian: unfamiliarity with idioms and reliance on fixed-definition replacements that ignore context. GPT-3.5's limited proficiency in Persian is especially evident in idiom translations, where many idioms are missed or misinterpreted. Table 7 illustrates these challenges for Fa→En translations.

All LLMs show higher BLEU and BERT scores for Fa→En compared to En→Fa translations, likely due to the models' stronger understanding of English. This makes them more proficient at converting Persian sentence structures into English. Our manual evaluation of GPT-3.5 outputs also indicates that Fa→En translations are more literal, contributing to the higher scores. As with En→Fa, the second and third prompts improve GPT-4o scores across all models.

Google Translate falls behind significantly in idiom translation and unlike the En→Fa results, combining it with LLMs does not generally improve model performance, as most models already achieve higher BERTScores than Google Translate.

6 Conclusion

We introduced two parallel datasets for Fa→En and En→Fa translation. The Persian idioms were sampled from our PersianIdioms resource, compiled from an online dictionary with 2,200 idioms and their meanings. Using these datasets, we evaluated multiple LLMs, Google Translate, and their combination, focusing on idiom translation accuracy and fluency. Our results show that Claude-3.5-Sonnet performs best in both directions. We also found that models generally translate English idioms more effectively than Persian ones. Performance varies by translation direction—for En→Fa, combining Google Translate and GPT-3.5 often yields the best results, while for Fa→En, single prompting outperforms combinations. Additionally, we evaluated existing automatic metrics and GPT-4o as a judge, confirming GPT-4o's reliability for assessing idiom translation accuracy.

7 Limitations

Our work is limited in several aspects, which we briefly discuss here.

- our parallel dataset contains only 200 examples for each direction for Persian and English translations. Increasing the size is definitely helpful in strengthening the quality of this dataset and our results.
- We focus only on Persian and English languages. Extending to other languages would be helpful to understand if some of our observations are general or not.
- We evaluate Google Translate as a strong NMT model here. Testing more diverse and especially open-source models could strengthen our arguments.

References

Anthropic. 2024. Claude 3.5 sonnet model card addendum.

Antonio Castaldo, Johanna Monti, et al. 2024. Prompting large language models for idiomatic translation. In *Proceedings of the First Workshop on Creative-text Translation and Technology*, pages 37–44.

Cohere For AI. 2024. c4ai-command-r-plus (revision 432fac1).

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

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

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *Preprint*, arXiv:2301.08745.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Akbar Karimi, Ebrahim Ansari, and Bahram Sadeghi Bigham. 2019. Extracting an english-persian parallel corpus from comparable corpora. *Preprint*, arXiv:1711.00681.

Omid Kashefi. 2020. Mizan: A large persian-english parallel corpus. *Preprint*, arXiv:1801.02107.

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.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *arXiv preprint arXiv:2309.08591*. Cs.CL.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume NAACL, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. *Preprint*, arXiv:2301.13294.

Makoto Nagao. 1984. A framework of a mechanical translation between japanese and english by analogy principle. In *Proc. of the International NATO Symposium on Artificial and Human Intelligence*, page 173–180, USA. Elsevier North-Holland, Inc.

OpenAI. 2023. Introducing chatgpt. Available: https://openai.com/blog/chatgpt.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko,

Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *Preprint*, arXiv:2009.09025.

Giancarlo Salton, Robert Ross, and John Kelleher. 2014. Evaluation of a substitution method for idiom transformation in statistical machine translation. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 38–42, Gothenburg, Sweden. Association for Computational Linguistics.

Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions. *Preprint*, arXiv:2006.09479.

David Stap, Eva Hasler, Bill Byrne, Christof Monz, and Ke Tran. 2024. The fine-tuning paradox: Boosting translation quality without sacrificing llm abilities. *Preprint*, arXiv:2405.20089.

Kenan Tang, Peiyang Song, Yao Qin, and Xifeng Yan. 2024. Creative and context-aware translation of east asian idioms with gpt-4. *Preprint*, arXiv:2410.00988.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *Preprint*, arXiv:2406.08464.

Masaru Yamada. 2024. Optimizing machine translation through prompt engineering: An investigation into chatgpt's customizability. *Preprint*, arXiv:2308.01391.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yafei Zhu, Daisy Monika Lal, Sofiia Denysiuk, and Ruslan Mitkov. 2024. From neural machine translation to large language models: Analysing translation quality of chinese idioms. *Proceedings of the International Conference on New Trends in Translation and Technology Conference 2024*.