Advancing Idiomatic Understanding: Evaluating GPT-3.5 and Google Translate for Persian-English Translations

Anonymous ACL submission

Abstract

Figurative language, especially idiomatic expressions, poses significant translation challenges due to its cultural and contextual nuances. Large Language Models (LLMs) like GPT-3.5 have shown greater capability in translating figurative language compared to state-of-the-art neural machine translation (NMT) systems. However, the impact of different prompting methods and combining NMTs and LLMs on idiom translation remains unexplored. This paper introduces two parallel datasets for Persian→English and English→Persian translation to address these challenges. The Persian idiom examples are sampled from our Persian-Idioms resource, which is compiled from an online dictionary and contains 2200 idioms with their meanings and popularity scores. Using these datasets, we evaluate GPT models, Google Translate, and their combination, focusing on idiom translation accuracy, fluency, and contextual relevance. Additionally, we assess existing automatic evaluation metrics and GPT-3.5 and GPT-4 for evaluating idiomatic translations. Our results indicate that while Google Translate shows superior fluency, GPT-3.5 excels in accurately translating idioms. We also show that models are better at translating English idioms than Persian ones, and different configurations of models perform differently depending on the direction of translation. We will release all our resources and annotations upon publication.

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 natural language processing (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 models struggle with translating idioms, often rendering them literally (Raunak et al., 2023; Dankers et al., 2022). Early machine translation efforts attempted to address this problem using idiom dictionaries or direct substitution (Salton et al., 2014; Nagao, 1984). However, new idioms continually emerge, their meanings can vary by context, and even identical meanings can result in different translations in the target language, 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."

Liu et al. (2023b) introduce two techniques to enhance the performance of transformer-based models in idiom translation. However, the advent of Generative Pre-trained Transformer (GPT) models (Brown et al., 2020) has enabled improvements in idiom translation without additional modifications to existing models. GPT models can achieve near state-of-the-art translation performance with few-shot prompting. Furthermore, due to their higher tendency towards non-literalness (Raunak et al., 2023) and greater paraphrastic capability (Hendy et al., 2023), their performance surpasses that of NMT models when dealing with figurative language (Raunak et al., 2023).

In this research, we focus on English and Persian and study the impact of various prompting methods on the quality of translations of sentences containing idiomatic expressions of these two languages. We also examine whether the combination of NMT models like Google Translate and an LLM yields better performance in idiom translation.

We first introduce a comprehensive resource for idiomatic expressions in Persian. This resource

captures idiomatic expressions and their meanings, including contextual usage examples. Using this dataset, we aim to bridge the significant gap in resources available for bilingual idiom translation and facilitate the development of more culturally aware language models.

Additionally, we produce two parallel English→Persian and Persian→English datasets consisting of sentences containing English and Persian idiomatic expressions. We then use these datasets to evaluate the performance of GPT-3.5, Google Translate, and a combination of these models in idiom translation across various settings.

Since the evaluation of idiom translation is challenging, we introduce novel evaluation metrics for translation quality and explore suitable replacements for the manual evaluation by calculating the correlation between existing automatic evaluation metrics and manually obtained scores. Furthermore, we assess whether GPT-4 and GPT-3.5 could be suitable replacements for human evaluation.

In summary, our main contributions are as follows. (i) A new resource for Persian idioms, called PersianIdioms, which includes about 2300 idioms, with their meaning, and scored popularity. A subset of 700 contains also example usage. Based on our knowledge, such a resource does not exist for the Persian language. (ii) A new parallel Persian→English and English→Persian datasets (100 examples each) with inputs containing at least one idiom. (Persian sentences are from PersianIdioms, and English ones are from various sources including EPIE (), Abadis.) (iii) Evaluating different prompting methods and also a combination of GPT-3.5 and Google-translate on idiom translation and showing how different they work for Persian→English and English→Persian.

2 Related work

We first review some of the related datasets for idiom expressions, followed by referencing a few recent work focusing on LLMs for translation.

2.1 Idiom datasets

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. (2023a) 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. (2023) present a methodology for constructing a large-scale, multilingual idiom knowledge base (IDIOMKB) 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. There remains a need for a large-scale Persian-specific idiom dataset. This research applies similar techniques of utilizing existing resources and language model generation to create idiom data specifically for Persian.

2.2 Translation using 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 own approaches and methods of translation. 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.

Despite these efforts, no study has systematically compared the performance of GPT models using different prompts, prompting methods, and combinations of GPT with NMT models. This work aims to address these gaps.

3 Datasets

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

3.1 PersianIdioms

Here, we explain how we build PersianIdioms. Our data collection beigns with extracting Persian idioms and their meanings from an online dictionary called Abadis[1]. For each idiom, we also gathered usage examples to provide contextual clarity, sourced from user-generated examples in Abadis. These examples are crucial for future testing of language models, allowing them to learn from actual idiomatic expressions in use.

To quantify the relevance of each idiom within a cultural context, we calculated a popularity score based on the number of pages each idiom has on Google. This score reflects its prevalence and cultural significance. Notably, this comprehensive dataset of Persian idioms, their meanings, contextual usage examples, and popularity scores has never existed before, making it a valuable resource for the development and evaluation of language models for Persian.

Data verification Once collected, the idioms are annotated with their meanings, examples, and popularity scores. This annotated data underwent a meticulous cleaning process to ensure accuracy and consistency. Subsequently, native Persian speakers reviewed the annotated idioms to verify the accuracy of meanings and the appropriateness of contextual examples. This manual verification process was critical to maintaining the linguistic integrity and cultural relevance of the dataset.

The culmination of these efforts is a dataset that contains 2,200 pairs of idioms and their corresponding meanings, annotated with popularity scores and supplemented with contextual examples where applicable that include 700 of them. This dataset is not only a testament to the richness of Persian idiomatic expressions but also a robust tool for advancing NLP capabilities in interpreting culturally nuanced language.

[1]https://abadis.ir/

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.
Popularity-Score	16800

Table 1: Persian idiom details in dataset, this popularity score shows that there are 16800 pages which contain this idiom in the internet.

3.2 Translation datasets

Persian→English From the idioms that contain contextual examples in the Persian idiom dataset, we select the top hundred with the highest popularity scores to perform idiom translation. Subsequently, a proficient translator produces English interpretations of these selected sentences, which are subjected to review and validation by another qualified expert. You can see an example of our data in Table 1.

English→Persian In our initial data collection phase, we tried to identify sentences containing idiomatic expressions from existing parallel English→Persian resources. However, upon examination, we discovered that the Persian translations within these datasets were either automatically generated, sourced from translations of English literature into Persian (Kashefi, 2020), or sourced from Wikipedia (Karimi et al., 2019). All of the mentioned approaches presented significant drawbacks to our research objectives. Automatic translation, such as that provided by models like Google Translate, often yields inaccurate results, particularly with figurative language, a phenomenon we aim to scrutinize in this study. Meanwhile, translations derived from English literature frequently incorporate contextual references, like character names, from other text sections, or alter sentence structures to

enhance fluency in the target language. Sentences extracted from Wikipedia tend to lack challenging, culturally-specific idioms and predominantly feature easily translatable expressions like "under pressure". Faced with these impediments, we opted for manual data collection. Primarily drawing from the EPIE corpus, we carefully selected sentences designed to spotlight the translation challenges posed by idiomatic expressions rather than the overall sentence structure. Subsequently, a proficient translator produced Persian renditions of these selected sentences, which were subjected to review and validation by another qualified expert.

The culmination of these endeavors is a dataset comprising 100 pairs of English sentences and their corresponding Persian translations. Given the existence of datasets containing English idioms and their meanings, we refrained from duplicating efforts in this regard.

4 Translation

In this section, we outline the models and settings used in our translation experiments and introduce the metrics we used to evaluate translation quality.

4.1 Methodology

We use Google Translate, GPT-3.5-turbo, and a hybrid approach combining both models to generate translations. For GPT-3.5-turbo, we experiment with three prompts, prompt chaining, and breaking down a single prompt into multiple steps, each used independently. More precisely, in prompt chaining, the model relies on chat history to generate an output. However, breaking down a single prompt into multiple independent prompts eliminates the need to rely on chat history. Based on our initial manual evaluation results for GPT3.5 and Google translate, described in Section 4.2, in the hybrid approach we employ GPT3.5 to identify and replace idioms with literal expressions, followed by using Google Translate to translate the resulting text into the target language.

The prompts used for GPT3.5-turbo for English→Persian translation are shown in Table 2. The second single prompt is taken from the prompts presented in Yamada (2024). Persian→English prompts replace "English" with "Persian" and vice versa, and "American" with "Iranian".

Based on the discussed prompting methods and models, we include these five different settings: (i) GPT3.5-turbo, Single Prompt (ii) GPT3.5-turbo, Prompt Chain (iii) GPT3.5-turbo, Multiple Prompts (iv) GPT3.5-turbo+Google Translate (v) Google translate.

4.2 Evaluation metrics

To evaluate the quality of translated text we first used BLUE (Papineni et al., 2002), BERT (Zhang et al., 2020), and COMET (Rei et al., 2020). However, upon manual inspection of the obtained results, we speculated that BLEU and BERT Scores are unsuitable indicators of a model's ability to identify and translate metaphors. The translations generated by Google Translate score the highest using both metrics. However, Google Translate demonstrates the weakest performance in idiom translation, often translating idioms word-for-word. Combining the MQM evaluation framework (Lommel et al., 2014) with this observation, we devise two mutually independent evaluation metrics, **fluency** and **idiom translation**. Idiom translation, which is either 0 or 1, focuses on whether the idiom is correctly translated in the context of the given sentence. Fluency, an integer between 1 and 5, focuses on the syntactic and semantic correctness of the translation, assuming the idiom is correctly translated. It is important to note that idiom translation concentrates only on the semantic accuracy of the generated translation. This means that if the idiom is correctly translated but contains grammatical errors, that error affects the "fluency" metric rather than the "idiom translation metric." Finally, the model's fluency is determined by averaging the fluency scores assigned to each of the 100 translations it produces, while idiom translation accuracy is measured by the percentage of idioms correctly translated.

In this new evaluation method, idiom translation replaces the "adequacy" metric in the MQM framework. We distill adequacy down to idiom translation for multiple reasons. First, the adequacy of translation relies heavily on the semantic correctness of the translated idiom. Second, our dataset consists of single sentences, that shift the translation challenge to the idiom itself. Therefore, if the idiom is correctly translated, the rest of the translation is likely to be semantically correct. Moreover, defining idiom translation as an independent subcategory of adequacy as outlined by MQM wouldn't be possible since the incorrect translation of an idiom often leads to semantic inaccuracies in other parts of the translation. Models frequently alter the original meaning of the sentence to produce a more

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.
Chain Prompts	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	1) Identify the idioms in this sentence and replace them with literal clauses. 2) Translate the literal sentence to Persian. Avoid word-for-word translation.

Table 2: Translation prompts used in our experiments.

coherent final translation. Consider the following example: Persian Sentence: گرسنه ام و میخواهم برم جیگرکی خودم رو بسازم. Gold translation is: "I'm hungry and I want to go to the liver shop and treat myself.". Model output is "I'm hungry and I want to make myself out of my liver." This example demonstrates how a semantic error in translating an idiom can lead to semantic errors outside the score of the idiom. Here, the Persian idiom "خودم رو بسازم" is translated literally to "make myself" instead of "treat myself", which has resulted in the incorrect translation of liver shop, ultimately changing the meaning of the whole sentence.

Lastly, idioms are the main focus of this work, and we aim to improve the performance of existing models in idiom translation while maintaining their performance in other aspects. Consequently, occasional semantic errors that are not due to idiom mistranslation affect the fluency score.

4.3 Automatic evaluation

Our experiments involve testing different prompting methods with varying numbers of shots across different models, resulting in a total of 27 configurations. Manually scoring them would be a tedious process. Consequently, we manually score five randomly selected configurations, each representing a different setting, across idiom translation and fluency. Then, we calculate the correlation between the scores given to each configuration and those obtained from the automatic evaluation metrics listed below. Metrics demonstrating acceptable correlation are subsequently used to evaluate the remaining outputs.

Existing automatic evaluation metrics In Section 4.2, we discuss that BLEU and BERT Scores are not suitable indicators of a model's idiom translation capabilities. In this section, we verify this hypothesis and evaluate whether these metrics and COMET could serve as suitable replacements for assessing fluency.

GPT4o Apart from the three automatic evaluation metrics initially chosen in Section 4.2, we use GPT-4o with the following prompt: "Are the idioms in this sentence correctly translated into Persian? Answer with just a number: 1 for yes and 0 for no." We also provide three examples to highlight the importance of correct idiom translation for the accuracy of the translations.

NLI Using GPT3.5 We employ natural language inference (NLI) to evaluate the performance of translation systems. Specifically, we prompt GPT-3.5-turbo to determine whether the model translations entail or contradict the gold translations. This method is validated by comparing the model-generated labels with manually assigned gold labels. Gold labels are derived from manual scores for idiom translation and fluency. A translation that scores "1" for idiom translation and exceeds the average fluency score is classified as an entailment; otherwise, it is classified as a contradiction.

We also calculate the correlation between the number of entailments for each configuration labeled by GPT-3.5 and the number of entailments labeled using the previously described criteria. While the model's assessment of entailment and contradiction might not perfectly align with human assessment, the number of entailments could serve as a reliable indicator of a model's performance relative to other models.

Results The first two rows of Table 4 display the correlation between manual scores for fluency and idiom translation metrics for the five selected configurations and BERT Score, COMET, BLEU, the number of correct translations according to GPT-4o, and the number of entailments according to GPT-3.5. The correlations between the idiom translation and BLEU indicate that BLEU penalizes non-literalness more than other automatic metrics.

BERT Score and COMET exhibit a similar, albeit less strong, tendency. Therefore, a lower BERT and BLEU score might indicate a higher quality of idiom translation. Ultimately, the high correlation between BERT Score and the fluency metric, and the number of translations labeled as correct by GPT-4o suggest that BERT Score and GPT-4o are reliable predictors of human scores for fluency and idiom translation from English→Persian, respectively.

The second two rows display the correlations for Persian→English translation. Similar to the English→Persian translations, the BLEU score penalizes non-literalness, and the number of correct translations identified by GPT-4o is a suitable replacement for idiom translation, highlighting the model's capability in idiom detection and understanding. Interestingly, COMET and BERT scores exhibit a higher correlation with idiom translation compared to fluency. This might suggest that translation from Persian→English is more likely to result in paraphrases and changes in sentence structure than English→Persian. However, these correlations are not high or low enough to draw definitive conclusions.

Table 3 shows the agreement percentage between the model's labels and manual labels in English→Persian translation. The agreement percentages between the model's labels and manual labels in English→Persian translation indicate that GPT-3.5 struggles to determine whether two sentences convey the same meaning. Moreover, these percentages are notably lower for Persian translations of English sentences, suggesting that GPT-3.5 models are less proficient in Persian compared to English. Consequently, GPT-3.5 may fail to recognize similar meanings between sentences that use different words in Persian.

Models	Agreement	
	Eng-Fa	Fa-Eng
GPT3.5, Single	56	62
GPT3.5, chain	58	73
GPT3.5, Multi	66	76
GPT3.5 + GT	55	64
GT	64	69

Table 3: Agreement percentage between the model's labels and manual labels in the NLI task for English→Persian and Persian→English translations. Setting names are shortened to fit the results in a single table. "GT" stands for "Google Translate".

5 Results

5.1 English→Persian

Table 5 demonstrates the results of the automatic and manual evaluation for models with the highest-performing hyper-parameters in each setting. If we denote the number of correct translations detected by GPT4 as "correct", in each setting, the best-performing hyper-parameters were chosen using the following formula: $modelscore = \frac{correct/100 + BERTScore}{2}$. The manual evaluation indicates that while GPT-3.5-turbo excels in understanding and translating idioms, Google Translate achieves higher fluency, as confirmed by the automatic evaluation metrics. As expected, the hybrid model, which integrates the strengths of both models, performs well in both fluency and idiom translation.

In the single prompt setting, the second prompt increases the number of correct translations by five and consistently performs better than the other two prompts with any number of shots. However, the number of shots does not significantly impact model performance in any setting, changing the results by only one or two percent. Given the limited number of sentences and the nondeterministic nature of GPT-3.5-turbo outputs, this effect could be coincidental. Therefore, we choose not to include these results.

Furthermore, it is evident from the results that metrics that rely on n-grams for evaluation like BLEU, are not suitable for evaluating translations that include figurative speech. Figurative speech can be rephrased in various ways that convey the same meaning. However, metrics that rely on n-grams disregard the semantic similarity of paraphrases and measure how close the generated translations are to human translation. This explains why the BLEU Score remains low despite other metrics showing a comparatively higher performance.

Table 7 illustrates the common strengths and shortcomings of each translation method for English to Farsi translation. Google Translate generates fluent translations but often translates idioms literally. Single prompts correctly detect and define idioms, but their definitions lack fluency and sound unnatural. Chain and multiple prompts produce more natural translations but still struggle with fluency. Finally, the combination of GPT-3.5 and Google Translate achieves the most natural and fluent results, accurately capturing the idiom's meaning in the given context and producing the most

Direction	metric	COMET	BERT Score	BLEU	GPT-4o	NLI
English→Persian	**Fluency**	74.36	91.96	92.88	-39.23	-59.45
	Idiom Translation	16.87	-56.88	-71.16	98.92	84.03
Persian→English	**Fluency**	50.35	32.43	50.67	11.84	37.62
	Idiom Translation	66.00	45.06	-15.41	85.79	46.19

Table 4: Correlation between results obtained from automatic evaluation metrics and results obtained manually for English→Persian and Persian→English translations.

	COMET	BERT Score	BLEU	GPT4o	GPT3.5 NLI	Idiom Transl.	Fluency
GPT3.5 Single Prompt	84.60	83.41	14.50	61	47	71	4.2
GPT3.5, Prompt Chain	84.21	82.17	9.74	67	59	82	4.22
GPT3.5, Multiple Prompts	85.42	83.24	12.26	79	**70**	91	4.25
GPT3.5 + Google Translate	**86.98**	84.35	16.53	**84**	49	**97**	4.58
Google Translate	82.77	**84.66**	**21.83**	42	37	42	**4.64**

Table 5: Results for automatic and manual evaluation of different translation models and prompts for English→Persian. "GPT3.5 NLI" shows the results for extrinsic evaluation using GPT3.5-turbo. Using a combination of GPT3.5-turbo and Google Translate not only brings their best qualities together but also increases their performance in both idiom translation and fluency.

coherent translations.

5.2 Persian→English

Table 6 presents the results of both automatic and manual evaluations for models with the highest-performing hyper-parameters in each setting. Due to the lack of a suitable automatic metric for fluency in Persian→English translation, we rely on the number of correct translations according to GPT-4o to select the best-performing hyper-parameters.

Manual evaluation indicates that similar to English→Persian, GPT-3.5 outperforms Google Translate in idiom translation. However, the scores given by GPT-4o and the idiom translation scores are significantly lower than their counterparts in English→Persian. Notably, the best performance in idiom translation is achieved not in the hybrid setting but in the GPT-3.5 single prompt setting. In this setting, the third prompt (see Table 2) consistently shows better results across all metrics compared to the other two prompts.

In Persian→English translation, the replacement process of idioms with literal expressions should be conducted in Persian. However, GPT-3.5 is relatively less fluent in Persian compared to English. Our manual evaluation identifies two major obstacles in the idiom replacement process in Persian: unfamiliarity with Persian idioms and the use of fixed-definition replacements. GPT-3.5's lack of proficiency in the Persian language is evident, particularly in idiom translations, where many idioms are either not detected or incorrectly defined. The more critical issue is the incorrect translations of correctly detected and defined idioms. Idioms are often replaced with a fixed definition, disregarding sentence context, which as mentioned in Section 1 can easily result in incorrect translations. These replacements also tend to ignore sentence structure, significantly reducing translation fluency. This might also explain the superior performance of the third single prompt. The first prompt does not provide any further instructions beyond translating the sentence, resulting in translations that are too literal. Conversely, the second prompt hints at the replacement of language-specific expressions, leading to the aforementioned replacement issues. These issues are evident in the model's explanations for each translation. An example demonstrating the translation quality in different settings is shown in Table 8.

Additionally, according to both manual and automatic evaluation results, GPT-3.5's better understanding of English sentences makes it more proficient at converting Persian sentence structures to English compared to the reverse. Persian→English translations tend to be more literal and word-for-word, which explains their comparatively higher performance in BLEU and BERT Scores.

	COMET	BERT Score	BLEU	GPT4o	GPT3.5 NLI	Idiom Transl.	Fluency
GPT3.5, Single Prompt	**73.35**	**92.63**	18.69	**40**	28	**40**	3.68
GPT3.5, Prompt Chain	68.9	92.02	17.91	20	25	23	3.82
GPT3.5, Multi. Prompts	67.52	91.87	16.16	18	31	25	3.33
GPT3.5 + Google Translate	71.7	92.28	21.55	25	**35**	37	**4.25**
Google Translate	71.57	92.53	**25.74**	19	28	21	3.78

Table 6: Results for automatic and manual evaluation of translation from Persian→English. "GPT3.5 NLI" shows the results for extrinsic evaluation using GPT3.5-turbo.

Sentence	Poor Mrs has lots of children and they were driving her up the wall!
Gold Translation	خانم بیچاره بچه های زیادی دارد که او را آزرده خاطر میکردند!
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.
Single Prompt	خانم فلانی بچه های زیادی دارد که دارند به اعصابش میزنند. Some Mrs has lots of children that are hitting her nerves.
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.
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!
GPT3.5+Google Translate	خانم بیچاره بچه های زیادی دارد که او را اذیت می کردند! Poor mother had many children that are bothering her!

Table 7: An example of English→Farsi translations. Back-translations into English are included to demonstrate translation quality to non-Persian speakers.

Sentence	من اعتماد کردم و حرف دلم را به او زدم اما او به هر کس رسید همه را روی دایره ریخت
Gold translation	I trusted and spoke my heart to him, but he **revealed all my secrets to everyone he encountered.**
Google Translate	I trusted and spoke my heart to her, but she **threw everyone on the circle**
Single Prompt	I trusted him and opened up to him, but he **betrayed my trust by sharing everything with everyone.**
Chain Prompt	I trusted him and confided in him, but he **betrayed everyone and turned against all.**
Multi Prompt	I trusted him and spoke my heart out to him, but he **reached out to everyone and spread it on everyone's face.**
GPT3.5+Google Translate	I trusted and told her my heart, but she **reached out to everyone and revealed everything**

Table 8: An example of Farsi→English translations.

6 Conclusion

In this paper, we introduced two parallel datasets for Persian→English and English→Persian translation. The Persian idiom examples were sampled from our PersianIdioms resource, which is compiled from an online dictionary and contains 2200 idioms with their meanings and popularity scores. Using these datasets, we evaluated GPT models, Google Translate, and their combination, focusing on idiom translation accuracy, fluency, and contextual relevance. Additionally, we assessed existing automatic evaluation metrics and GPT-3.5 and GPT-4 for evaluating idiomatic translations. Our results indicated that while Google Translate shows superior fluency, GPT-3.5 excels in accurately translating idioms, especially for English→Persian.

We also showed that models are generally better at translating English idioms than Persian ones. Different configurations of models perform differently depending on the direction of translation. For example, in English→Persian, the combination of Google Translate and GPT-3.5 works the best, while for Persian→English, it is outperformed by a simple single prompting of GPT-3.5. Additionally, we provided a comprehensive analysis of the underlying reasons for this disparity. Further, our results suggest that strong multilingual LLMs such as GPT-4o can act as evaluators for the case of English and Persian idiom translations. This is significant as parallel data annotation is expensive and also shows how strong these models are even for challenging figurative language.

7 Limitations

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

- our parallel dataset contains only 100 examples for each direction for Persian and English

translations. The examples are mostly annotated using one expert. Increasing the size and the number of annotators is definitely helpful in strengthening the quality of this dataset.

- 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.

- Our tested LLMs are limited to only OpenAI GPT models, which are not open-source. While this has simplified our current experimentation, it is necessary to extend our evaluations to include more LLMs, especially open-source ones.

- We evaluate Google Translate as our NMT model here. Similar to LLMs, testing more diverse and especially open-source models could strengthen our arguments.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

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.

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. 2023. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. *arXiv preprint arXiv:2308.13961*. Cs.CL.

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

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023b. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15095–15111, Singapore. Association for Computational Linguistics.

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.

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 Awadalla. 2023. Do gpts produce less literal translations? *Preprint*, arXiv:2305.16806.

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.

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*.