Comparative Study of Multilingual Idioms and Similes in Large Language Models

Paria Khoshtab[*1] **Danial Namazifard**[*1]
Mostafa Masoudi[1, 2] **Ali Akhgary**[1] **Samin Mahdizadeh Sani**[1]
Yadollah Yaghoobzadeh[2, 1]
[1]School of Electrical and Computer Engineering
College of Engineering University of Tehran, Tehran, Iran
[2]Tehran Institute for Advanced Studies, Khatam University, Tehran, Iran
{paria.khoshtab, namazifard, mostafa.masoudi, ali.akhgary, samin.mehdizadeh, y.yaghoobzadeh}@ut.ac.ir

Abstract

This study addresses the gap in the literature concerning the comparative performance of LLMs in interpreting different types of figurative language across multiple languages. By evaluating LLMs using two multilingual datasets on simile and idiom interpretation, we explore the effectiveness of various prompt engineering strategies, including chain-of-thought, few-shot, and English translation prompts. We extend the language of these datasets to Persian as well by building two new evaluation sets. Our comprehensive assessment involves both closed-source (GPT-3.5, GPT-4o mini, Gemini 1.5), and open-source models (Llama 3.1, Qwen2), highlighting significant differences in performance across languages and figurative types. Our findings reveal that while prompt engineering methods are generally effective, their success varies by figurative type, language, and model. We also observe that open-source models struggle particularly with low-resource languages in similes. Additionally, idiom interpretation is nearing saturation for many languages, necessitating more challenging evaluations.[1]

1 Introduction

Large language models (LLMs) have revolutionized NLP by demonstrating remarkable capabilities in understanding and generating human language. One of the most challenging aspects of human language for LLMs to comprehend is figurative language, which includes similes, idioms, and metaphors. Figurative language significantly enriches human communication by facilitating the implicit expression of complex ideas and emotions (Roberts and Kreuz, 1994; Fussell and Moss, 2014). Unlike literal expressions, figurative language often involves rich cultural references and judgments that vary considerably across different cultures (Shutova, 2011; Fussell and Moss, 2014). Consequently, understanding and generating figurative language is crucial for LLMs to interact naturally and effectively with users. Therefore, studying how these models handle figurative language is essential for advancing their capabilities.

Recent studies have highlighted that LLMs, not only struggle to generate but also frequently misinterpret figurative expressions (Huang et al., 2024), underscoring the need for more sophisticated techniques to bridge these gaps. The challenge becomes even more pronounced in multilingual contexts, where figurative language is intricately tied to cultural nuances (Liu et al., 2024b).

There remains a gap in the literature regarding the comparative performance of LLMs in interpreting different types of figurative language, in English and multilingual contexts. This study focuses on two types of figurative language: similes and idioms. A simile compares two entities, typically using “like” or “as” (e.g., “as busy as a bee”), to create vivid descriptions. An idiom, by contrast, is a fixed phrase whose meaning cannot be inferred from the meanings of its components (e.g., “kick the bucket”). These two are distinct in their structure and usage, suggesting that LLMs might perform differently on them, and require different strategies to process them effectively.

We evaluate the performance of multiple LLMs across various languages using two existing datasets: MABL (Kabra et al., 2023), MAPS (Liu et al., 2024b), and our newly developed Persian datasets for simile and idiom interpretation. MABL includes examples of figurative language interpretation as an inference task, and mainly simile expressions. It covers eight languages in high and low resource ranges. MAPS is a multilingual dataset of proverb interpretation including six languages. To further advance this research, we contribute by extending the scope to Persian, the native language

[*] Equal contribution, ordered alphabetically.

[1] Data and code: https://github.com/namazifard/Multilingual-Idioms-Similes

Proceedings of the 31st International Conference on Computational Linguistics, pages 8680–8698
January 19–24, 2025. ©2025 Association for Computational Linguistics

of the authors, by developing two additional evaluation sets. These new sets help us analyze the datasets and model performance more deeply.

Evaluating LLMs requires interacting with them, making prompt engineering a critical component for optimizing performance. We examine various prompting strategies, including chain-of-thought, few-shot, and dialogue simulation. We extend our evaluations to input being translated to English, as it serves as a strong baseline for many multilingual evaluations (Lin et al., 2022; Shi et al., 2023; Liu et al., 2024b). To achieve a comprehensive evaluation, we conduct an exhaustive assessment using both closed-source—GPT-3.5, GPT-4o mini, and Gemini 1.5 (Team et al., 2024)—and open-source Llama 3.1 (Dubey et al., 2024, 8B, 70B) and Qwen2 (Yang et al., 2024, 7B, 72B).

Our findings reveal several novel insights: (i) Prompt engineering methods show varying degrees of success depending on the figurative type, language, and model used. (ii) Open-source models perform similarly to closed-source models in idioms, but they generally lag behind in interpreting similes. (iii) The interpretation of idioms in the style of the MAPS dataset is nearing saturation for many languages when strong LLMs are used, due to the presence of idioms and their meanings in their training data. (iv) The presentation of pre-trained data, as well as the script used in different languages, significantly impact model performance (v) Chain-of-thought prompting proves particularly effective for simile interpretation in smaller models.

2 Related work

2.1 Figurative language processing

Figurative expressions encapsulate complex human experiences and cultural knowledge, making them essential in tasks ranging from sentiment analysis (Hercig and Lenc, 2017) to machine translation (Wang et al., 2024a). Previous research efforts have focused on various types of figurative language, with several approaches dedicated to improving simile detection and component extraction (Qadir et al., 2015; Mpouli, 2017; Liu et al., 2018; Zeng et al., 2019), as well as on generating similes (Chakrabarty et al., 2020; Zhang et al., 2021; Lai and Nissim, 2022).

In addition to similes, proverbs and idioms are another type of figurative expression that has been explored in various studies, focusing on identifying whether a phrase is used idiomatically or proverbially, either within a specific context (token-level) or in general (type-level) (Li and Sporleder, 2009; Fazly et al., 2009; Verma and Vuppuluri, 2015; Salton et al., 2016; Peng and Feldman, 2016). Beyond detection and generation, researchers have examined methods for interpreting and representing these figurative expressions, including literal paraphrasing, treating them as single tokens, or composing them from characters rather than words (Liu and Hwa, 2016; Zhou et al., 2021). Additionally, significant research has been conducted on other types of figurative language, such as metaphors, sarcasm, and irony (Yu and Wan, 2019; Ghosh et al., 2017; Chakrabarty et al., 2021).

While many studies have focused on multilingual figurative language detection (Lai et al., 2023; Tedeschi et al., 2022; Tayyar Madabushi et al., 2022; Aghazadeh et al., 2022) and English multi-figurative (multiple types of figurative) language processing (Jhamtani et al., 2021; Chakrabarty et al., 2022), our research centers on multilingual multi-figurative language interpretation, which is a highly understudied area.

Our work expands upon the foundation laid by Liu et al. (2024b) by introducing both similes and idioms across a wider range of languages, including the creation of new Persian datasets for each figurative type. While Liu et al. (2024b) focus exclusively on proverbs and sayings, we extend the analysis to similes, offering a broader evaluation of LLMs' figurative language comprehension, and a comparative study of them.

2.2 Multilingual prompt engineering

While LLMs have achieved impressive success in various NLP tasks (Brown et al., 2020), they encounter significant challenges in tasks that require understanding culturally specific figurative language (Li et al., 2024b). The high cost of collecting multilingual cultural data further complicates these tasks. As a result, current methods to enhance the cultural awareness of LLMs rely primarily on two approaches: prompt engineering and culture-specific pre-training (Li et al., 2024a).

One prominent strategy is **Chain-of-Thought (CoT)** prompting, which has been demonstrated to improve LLM performance on various reasoning tasks by breaking down complex problems into more manageable steps (Wei et al., 2023). Shi et al. (2023) highlight the effectiveness of multilingual CoT prompting on reasoning benchmarks, though their evaluation does not include figurative language interpretation, such as similes and idioms, which is central to our work. We extend this research by applying CoT prompting specifically to simile interpretation, where cultural reasoning is often required. Another promising technique is **Reasoning in Conversation (RiC)**, introduced by Wang et al. (2024b), which simulates dialogue to improve performance on subjective, culturally-related tasks. In our study, we apply RiC specifically to simile tasks, leveraging the conversational context to enhance model reasoning for these culturally nuanced expressions.

In addition to prompt engineering, **translate-test** methods are commonly employed to address multilingual challenges. In this approach, evaluation data is translated into English before processing, using tools like Google Translate or LLMs (Ahuja et al., 2023; Liu et al., 2024a; Shi et al., 2023). This method has proven effective in reducing performance gaps across languages (Conneau et al., 2018; Ponti et al., 2020; Artetxe et al., 2023). Building on this, Huang et al. (2023) propose Cross-Lingual-Thought (XLT) prompting, that translates the question into English and solves the problem in English before generating a response in the original language. We study the impact of English translation on simile and idiom interpretation across several languages, evaluating the consistency of responses between original and translated inputs under zero-shot and CoT settings.

3 Methodology

We focus on the task of figurative language interpretation, specifically across two types of figurative expressions: similes and idioms. Additionally, we extend the evaluation to Persian by creating new test sets for this language. Below, we describe the datasets used in our study, followed by an overview of the LLMs and prompting techniques applied.

3.1 Datasets

We employ two existing datasets: MABL (Metaphors Across Borders and Languages) (Kabra et al., 2023) for simile experiments, and MAPS (Multicultural Proverbs and Sayings) (Liu et al., 2024b) to assess idioms. Both datasets facilitate the analysis of figurative language understanding across multiple languages, providing a diverse multilingual resource. A description of the datasets can be found in Appendix A; however, key points are explained in the following.

MABL contains figurative expressions in eight languages: English, Indonesian, Hindi, Swahili, Yoruba, Kannada, Sundanese, and Javanese. This dataset captures cultural and linguistic diversity in figurative language, offering a valuable resource for testing multilingual LLMs' abilities. We randomly select 200 simile samples (to be close to the number of examples in the idiom dataset, ensuring balance and comparability between the two datasets) from each language for evaluation, as shown in Table 1.

start phrase	ending 0	ending 1	label
The test is as easy as rocket science	The test is easy	The test is hard	1

Table 1: An English example from MABL.

MAPS consists of proverbs and sayings, designed to evaluate their interpretation within conversational contexts. The dataset provides binary labels, indicating whether the proverb is used figuratively. It spans six languages: English, Indonesian, Mandarin Chinese, Bengali, German, and Russian, with sample counts of 214, 267, 143, 272, 183, and 226 respectively. We select idiomatic sentences for evaluation, as detailed in Table 2.

proverb	conversation	answer A	answer B	label
fair exchange is no robbery	Person 1: "Can I borrow your pen?" Person 2: "Sure, can I borrow your notebook?" Person 1: "Fair exchange is no robbery"	Person 1 will lend the notebook to Person 2.	Person 1 will not lend the notebook to Person 2.	A

Table 2: An English example from MAPS.

Persian datasets In addition to using the existing datasets, we created two new datasets specifically for the Persian language, following the formats of MABL and MAPS for similes and idioms that provide resources for evaluating figurative language understanding in Persian, which is underrepresented in current multilingual model researches.

(i) Persian simile We follow the methodology used in the MABL dataset but utilize GPT-4o to assist in generating examples. The model generates simile pairs by producing a start phrase with two possible endings—one reflecting the correct simile interpretation and the other conveying an incorrect meaning. The Persian Simile dataset consists of 200 samples. The prompts closely follow the instructions from the MABL dataset, ensuring consistency with the format and objectives of the original dataset (Appendix B). After generation, three native Persian speakers manually evaluate and correct the examples to ensure both accuracy and cultural relevance.

(ii) Persian idiom We follow a methodology inspired by the creation process of the original MAPS dataset. First, we collect Persian idioms from two online resources: Daneshchi [2], an educational portal, and Abadis [3], an online dictionary. To create conversational contexts based on each idiom's explanation, we develop prompts using examples written by native Persian speakers to guide the model in understanding how idioms are used in everyday conversation. Using GPT-4o, we generate a conversational context for each idiom, ensuring that the idiom is correctly situated in a natural dialogue. In the second round, we provide the model with the idiom, explanation, and the generated conversational context. The model is then tasked with generating two response choices—one correct and one incorrect—based on the meaning of the idiom in the conversation. The Persian Idiom dataset contains 316 samples. Finally, native speakers review the generated content to ensure its accuracy, cultural appropriateness, and grammatical correctness.

For more details on the dataset construction and verification process, see Appendix B.

3.2 Language Categorization

Based on our experiment results we propose a categorization of languages from two perspectives, as shown in Figure 1. The script of a language plays a significant role, as Latin-based languages like English dominate the pretraining data of large language models. Additionally, Joshi et al. (2020) classify languages into six categories based on the availability of labeled and unlabeled data on the web. This classification, alongside the language script, offers valuable insights for analyzing model behavior.

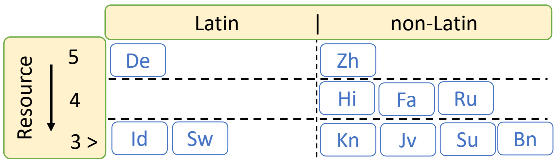


Figure 1: Our language categorization provides clearer insights for analyzing the results.

3.3 Models

We evaluate the performance of several open-source and closed-source LLMs to understand their capabilities in processing figurative language across multiple languages. The models under consideration include GPT-3.5-Turbo-0125 (OpenAI, 2023), GPT-4o mini (OpenAI et al., 2024), and Gemini 1.5 Flash (Team et al., 2024) which are representative of closed-source commercial LLMs, as well as open-source models: Llama 3.1 (Dubey et al., 2024, 8B, 70B) and Qwen2 (Yang et al., 2024, 7B, 72B). These models are selected based on their widespread use and the contrast they offer in terms of accessibility, customization potential, and model size variations. Note that although Llama 3.1 does not cover all of our languages (Just En, De, and Ru are covered), it performed quite well in our initial experiments. The cost for running experiments is given in Appendix C.

3.4 Prompting

We use several prompting techniques with examples in native or English translation or combined. We explain the techniques and then mention which ones we used in native, English, and native-English setups. The instruction is consistently given in English across all settings, as shown in Table 3. The full prompt templates used in the experiments are available in Appendix D.

3.4.1 Techniques

We explore various prompt engineering techniques: (i) **Zero-shot**: this setting assesses the model's basic understanding. (ii) **One-shot**: here, we explore how providing a single example can enhance the model's knowledge of the task and its cultural context. (iii) **Chain of Thought (CoT)**: this approach leverages the model's reasoning capabilities to break down and process figurative meanings step by step (Wei et al., 2023), which is still effective under multilingual scenarios (Shi et al., 2022). The instructions guide the model through a thought process to interpret idioms or similes. Additionally, a one-shot example is provided in the native language, accompanied by an explanation in the same language. For idioms, the one-shot example only explains the proverb's meaning. However, for similes, a reasoning pathway is provided that involves (1) mentioning the target culture, (2) interpreting the simile's meaning, (3) clarifying the reason for the similarity and its connection to the start phrase, and (4) generating the final answer. Examples of

[2] https://www.daneshchi.ir

[3] https://abadis.ir

CoT are shown in Table 3. (iv) **Dialogue simulation**: since figurative expressions often deviate from their literal meaning, understanding can be improved by placing them in context (Liu et al., 2024b). We use the RiC (Reasoning in Conversation) method (Wang et al., 2024b) in zero-shot settings to prompt the model to generate dialogues or conversations between two individuals, embedding the figurative expression within these interactions.

Simile CoT example of Swahili

In this task, you are given a start phrase indicating a figurative expression in Swahili culture. Please select 0 if the start phrase conveys the meaning of ending 0, and 1 if it conveys the meaning of ending 1.
Start phrase: He felt his chest was frozen like ice
Ending 0: His heartbeat went too fast
Ending 1: His heartbeat went slowly
Answer: In Kiswahili culture, the expression "He felt his chest was frozen like ice" is used to show that a person felt fear, anxiety or uncertainty. "The chest is frozen like ice" means that his emotions were stuck or frozen like ice. Therefore, the expression indicates a lack of enthusiasm. In this context, "His heartbeat was slow" indicates that the person's heartbeat was slow and steady. This indicates a sense of unhurriedness or excitement. Therefore, the answer is 1.
<Test Simile Example>

Idiom CoT example of German

Question: How would one interpret this proverb in German culture, given the context? Please first think about the proverb's meaning, then write an explanation of the proverb's meaning, and finally choose between A and B.
Proverb: hope dies last
Context: Person 1: Do you think we'll make it? Person 2: I don't know, but hope dies last.
Choices: A: Person 2 hopes they'll make it. B: Person 2 has no hope they'll make it.
Explanation: hope should be the last thing you give up
Answer: A
<Test Idiom Example>

Table 3: One-shot example of CoT prompts for simile and idiom. Thought pathways are specified with colors. blue: specify culture, cyan: analyzing expression meaning, orange: connect it to StartPhrase, green: specify the final answer. purple: additional CoT trigger for idioms.

3.4.2 Input language

Native In this setting, the input examples are presented in their original language (e.g., Chinese, Indonesian, etc.) along with instructions in English. We evaluate our experiments across three different configurations: zero-shot, one-shot, and CoT.

Translated-English We follow the trend of translating either the training or test data into English (Shi et al., 2023; Conneau et al., 2018; Qin et al., 2023), utilizing three translation systems that cover all of our languages: the Google Translate API, Meta's No Language Left Behind (NLLB-200-3.3B) model (Team et al., 2022), and GPT-3.5 (OpenAI, 2023). We investigate zero-shot and one-shot settings to explore the few-shot effect in this context. For CoT, we ask the model to explain the meanings of idioms or similes alongside a one-shot example in English to activate CoT reasoning abilities. For simile, we also experiment with dialogue simulation prompting techniques. For idiom tasks, as they are already presented in a dialogue format, dialogue simulation is not applied.

Native and Translated-English In this approach, we directly prompt the model to convert the input from the native language to English and then perform the task (Etxaniz et al., 2024; Huang et al., 2023). The input is provided in both native and Translated-English simultaneously, leveraging the model's intrinsic translation capabilities. This technique is applied within a one-shot setting, as detailed in Tables 11 and 12.

3.5 Evaluation Process

We utilize LLMs to evaluate simile and idiom tasks. For each task, we select an LLM, apply the appropriate prompt, and retrieve the model's answer. The answers are parsed using regular expressions (regex) to extract the final binary result. In cases where regex fails to correctly extract the answer due to irregular formatting or model output variations, manual verification is employed to ensure accuracy in the evaluation process.

4 Results

In this section, we present the performance of various large language models (LLMs) across two datasets and multiple languages.

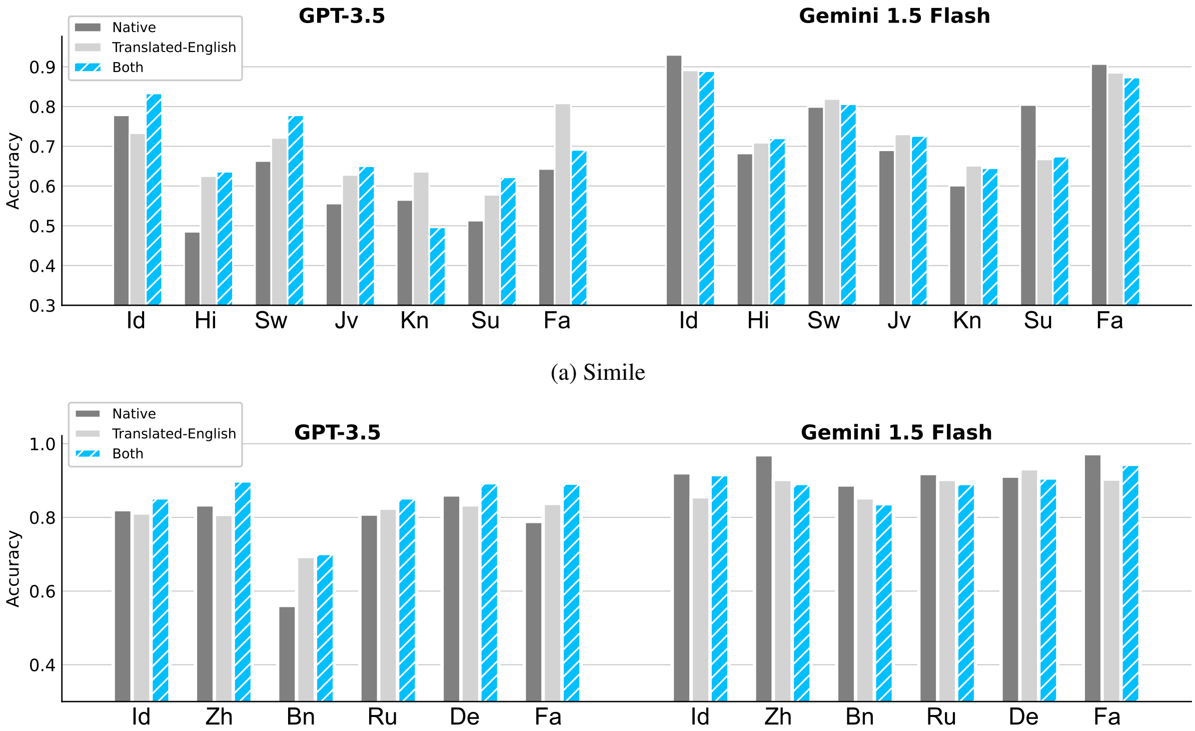
4.1 Native

We evaluate the performance of LLMs on input provided in native languages. Results for closed- and open-source models are shown in Table 4. The results show that both types of models can comprehend similes and idioms to varying extents. Generally, performance on idioms is higher than similes.

For similes, open-source models with fewer parameters show lower accuracies, with Llama 3.1-8B outperforming Qwen2-7B. Large open-source models, i.e., Llama 3.1-70B and Qwen2-72B, outperform GPT-3.5. Among all models, however, Gemini 1.5 demonstrates the best simile interpretation performance on average across our two prompt-

	Open Source								Closed Source					
Language	Qwen2-7B		Qwen2-72B		Llama 3.1 8B		Llama 3.1 70B		GPT-3.5		Gemini 1.5 Flash		GPT-4o mini	
	Zero-Shot	CoT	Zero-Shot	CoT	Zero-Shot	CoT	Zero-Shot	CoT	Zero-Shot	CoT	Zero-Shot	CoT	Zero-Shot	CoT
SIMILE														
En	.651$_{.010}$	.830$_{.010}$	.936$_{.008}$	**.943**$_{.006}$	.761$_{.030}$	.770$_{.010}$	.913$_{.008}$	.883$_{.011}$	.786$_{.006}$	.916$_{.010}$	.878$_{.011}$	.896$_{.006}$	.740$_{.014}$	.916$_{.008}$
Id	.590$_{.031}$	.621$_{.035}$	.893$_{.006}$	.911$_{.009}$	.643$_{.027}$	.728$_{.018}$	.890$_{.010}$	.895$_{.010}$	.683$_{.023}$	.805$_{.014}$	.910$_{.007}$	**.925**$_{.004}$	.711$_{.010}$	.911$_{.004}$
Hi	.530$_{.018}$	.583$_{.018}$	.635$_{.010}$	.638$_{.009}$	.533$_{.048}$	.566$_{.004}$	.655$_{.016}$	.691$_{.010}$	.531$_{.013}$	.615$_{.010}$	.676$_{.004}$	**.740**$_{.007}$	.606$_{.022}$	.685$_{.008}$
Sw	.508$_{.012}$	.498$_{.013}$	.618$_{.002}$	.596$_{.040}$	.520$_{.017}$	.523$_{.013}$	.693$_{.026}$	.681$_{.031}$	.533$_{.006}$	.714$_{.007}$	.745$_{.014}$	**.788**$_{.012}$	.556$_{.013}$	.768$_{.011}$
Jv	.501$_{.008}$	.536$_{.014}$	.573$_{.015}$	.656$_{.022}$	.501$_{.029}$	.568$_{.019}$	.669$_{.004}$	.641$_{.020}$	.490$_{.008}$	.571$_{.020}$	.698$_{.010}$	.673$_{.009}$	.606$_{.009}$	**.783**$_{.004}$
Kn	.491$_{.023}$	.493$_{.002}$	.465$_{.014}$	.514$_{.006}$	.403$_{.023}$	.488$_{.009}$	.493$_{.035}$	.610$_{.008}$	.561$_{.002}$	.530$_{.003}$	.588$_{.004}$	**.619**$_{.004}$	.525$_{.020}$	.576$_{.006}$
Su	.435$_{.022}$	.505$_{.020}$	.548$_{.025}$	.568$_{.006}$	.456$_{.037}$	.490$_{.008}$	.593$_{.023}$	.586$_{.048}$	.485$_{.004}$	.583$_{.004}$	.746$_{.010}$	.753$_{.009}$	.590$_{.014}$	**.768**$_{.012}$
Fa	.576$_{.013}$	.623$_{.009}$	.855$_{.010}$	.864$_{.008}$	.658$_{.024}$	.810$_{.020}$	.851$_{.014}$	**.926**$_{.015}$	.600$_{.017}$	.773$_{.006}$	.915$_{.004}$	.830$_{.018}$	.680$_{.010}$	.898$_{.002}$
Average	.535	.586	.690	.711	.559	.618	.719	.739	.583	.688	.769	.778	.627	**.788**
IDIOM														
En	.925$_{.003}$	.915$_{.004}$	.981$_{.002}$	**.990**$_{.004}$	.878$_{.008}$	.887$_{.004}$	.975$_{.002}$	.982$_{.004}$	.970$_{.004}$	.970$_{.002}$	.953$_{.000}$	.948$_{.006}$	.976$_{.005}$	.981$_{.004}$
Id	.853$_{.008}$	.831$_{.011}$	.917$_{.003}$	**.920**$_{.004}$	.711$_{.007}$	.801$_{.002}$	.895$_{.015}$	.914$_{.008}$	.852$_{.007}$	.789$_{.009}$	.900$_{.004}$	.912$_{.009}$	.918$_{.008}$	.894$_{.007}$
Zh	.878$_{.009}$	.871$_{.006}$	.979$_{.003}$	**.986**$_{.003}$	.815$_{.010}$	.767$_{.008}$	.942$_{.003}$	.953$_{.009}$	.874$_{.000}$	.920$_{.014}$	.979$_{.000}$	.951$_{.009}$	.921$_{.006}$	.944$_{.011}$
Bn	.659$_{.006}$	.643$_{.004}$	.896$_{.002}$	**.913**$_{.004}$	.663$_{.010}$	.649$_{.008}$	.838$_{.016}$	.855$_{.014}$	.510$_{.008}$	.611$_{.009}$	.861$_{.001}$	.849$_{.009}$	.845$_{.010}$	.872$_{.007}$
Ru	.845$_{.003}$	.823$_{.004}$	.933$_{.002}$	**.938**$_{.006}$	.778$_{.007}$	.805$_{.012}$	.914$_{.014}$	.917$_{.004}$	.838$_{.005}$	.849$_{.009}$	.914$_{.007}$	.910$_{.007}$	.877$_{.004}$	.874$_{.004}$
De	.862$_{.002}$	.830$_{.006}$	.939$_{.003}$	.950$_{.004}$	.759$_{.009}$	.803$_{.006}$	.943$_{.003}$	**.957**$_{.004}$	.877$_{.002}$	.901$_{.004}$	.894$_{.009}$	.878$_{.005}$	.901$_{.002}$	.911$_{.006}$
Fa	.841$_{.012}$	.788$_{.010}$	.943$_{.010}$	.955$_{.007}$	.756$_{.005}$	.813$_{.009}$	.924$_{.018}$	.941$_{.012}$	.828$_{.004}$	.853$_{.010}$	**.964**$_{.001}$	.943$_{.002}$	.940$_{.009}$	.936$_{.009}$
Average	.837	.814	.941	**.950**	.765	.789	.918	.931	.821	.841	.923	.913	.911	.916

Table 4: Results of open-source and closed-source LLMs with native input languages, averaged over three runs (std is also shown). The best accuracy for each category of models is underlined, and the best overall accuracy is **bold**.



(a) Simile

(b) Idiom

Figure 2: Comparing results of inputs being in native, Translated-English, and both languages in one-shot setting.

ing strategies. Notably, Gemini 1.5 significantly boosts zero-shot performance in Sundanese (a low-resource language), achieving up to 73% accuracy, while the second-best model remains below 60%.

For idioms, the open-source model Qwen2-72B outperforms other models, even the closed-source ones. Interestingly, Qwen2, with only 7B parameters, outperforms GPT-3.5 in zero-shot on average, especially in Bengali and Persian. Among closed-source models, Gemini 1.5 remains the best. Notably, Gemini 1.5 demonstrates significant performance in Chinese and Persian, achieving accuracy levels even higher than in English, which highlights its strong capability in handling these specific languages compared with other models.

Overall, the results reveal that model performance varies significantly depending on the type of figurative language being evaluated. Moreover, they highlight that in understanding figurative language, an open-source model can outperform closed-source ones in specific languages. This suggests that success in comprehending figurative expressions may depend more on specific language or cultural knowledge encoded in the models, which might need further exploration in future work.

Impact of model size and CoT Since we do not have access to the size of closed-source models, we consider only open-source models when analyzing the effect of model size here. We observe that in nearly all cases, increasing the model size improves accuracy for both simile and idiom tasks. The effect is particularly pronounced in simile interpretation, where performance improves by about 16% absolute point on average. For specific languages such as English and Indonesian, the improvement reaches around 30%. In contrast, the improvement for idiom tasks is less significant on average.

Our analysis reveals that the effectiveness of **CoT** prompting varies by model size. For similes, smaller models, such as Qwen2-7B and Llama 3.1 8B, exhibit greater error reduction with CoT prompting, with reductions of approximately 11% and 13%, respectively. In comparison, larger models like Qwen2-72B and Llama 3.1 70B show smaller but still notable reductions, each around 7%. This suggests that CoT is particularly beneficial for smaller models in simile interpretation, likely by enhancing their reasoning capabilities.

For idioms, however, the pattern is different. While Llama 3.1 8B achieves a modest error reduction of 10%, Qwen2-7B experiences a performance drop of 14%, indicating that smaller models may struggle with the complexity of idiom interpretation when CoT is applied. Conversely, larger models like Qwen2-72B and Llama 3.1 70B demonstrate strong error reductions of over 15% each. These findings suggest that while larger models benefit from CoT across both tasks, their already strong reasoning capabilities make CoT less impactful for simile interpretation but more crucial for idiomatic expressions. Similar trends have been observed in other studies, such as Sprague et al. (2024), where the impact of CoT was found to be more pronounced in smaller open-source models than in larger ones.

On the other hand for closed-source models, CoT compared to zero-shot has improved efficiency for GPT-4o mini and GPT-3.5, though this trend is not consistently observed for Gemini 1.5. Additionally, CoT tends to be more effective for similes.

Cross-lingual interference in smaller models During our experiments, we observe that smaller models, particularly Llama 3.1 8B, occasionally exhibit cross-lingual interference. For instance, while generating responses in Persian, the model sometimes inserts Chinese characters or words in the middle of the response, only to revert back to Persian in the continuation of the text. This inconsistency suggests that smaller models may struggle to maintain coherence in the intended language, potentially due to confusion in multilingual settings.

4.2 English translation

Figure 2 represents our evaluations on GPT-3.5 as our base model and Gemini 1.5 Flash as the strongest closed-source model. We experiment with native and Translated-English inputs in the one-shot setting. Google Translate is used for our translation. Other translation methods are examined in Section 4.2.1

Comparing Native with Translated-English In both figurative tasks, GPT-3.5 exhibits lower accuracy across all languages when compared to Gemini 1.5, a trend also evident in native prompting. The use of translation has made a significant improvement in the performance of GPT-3.5 for similes. We can also see this trend with Gemini 1.5, but it is not consistent for all languages. With Gemini 1.5, native prompting surpasses translation for Indonesian, Persian, and Sundanese, which may show the superiority of the Gemini model in understanding these languages in their native format. When it comes to idioms, Gemini 1.5 performs slightly better in native prompts, though the improvement is minor. GPT-3.5, however, shows varying results depending on the language. For lower resource and non-Latin languages like Bengali, Persian, and Russian, GPT-3.5 tends to achieve better results when using Translated-English rather than native prompts, likely due to its weaknesses in handling these languages. Generally, it can be concluded that translation efficiency depends on the language, LLM, and task. This observation aligns with the conclusions drawn in the study by Zhang et al. (2023), suggesting that translation's effectiveness depends on whether potential comprehension gains outweigh translation errors, and it may not always enhance performance.

Input in Native and Translated-English The results of using both native and English inputs (as explained in Section 3.4.2) are shown in Figure 2 in the third column named "Both". When the GPT-3.5 model is used, using both native and Translated-English is a more effective method in most languages in both figurative types. There are two exceptions in similes for Kannada and Persian, where the Translated-English approach performs better. This seems to be specific to certain languages and requires further investigation. When using Gemini 1.5, a more capable model, the results are similar to those obtained using the Translated-English method for both figurative types. We also observe that Gemini 1.5 generally doesn't outperform native prompting in idioms. This indicates the model's strong understanding of idioms in native, and translating them may lead to errors in interpretation.

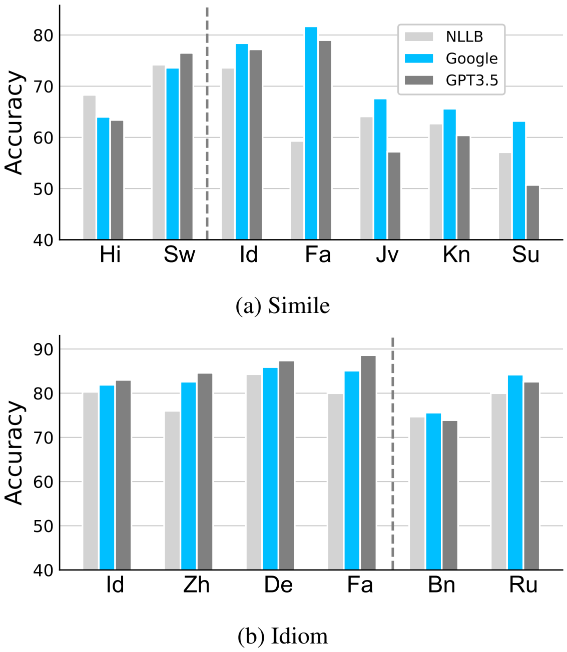
4.2.1 Comparing translation methods

So far in the reported results, the translation is done using only Google Translate. Here, we investigate two additional methods (NLLB and GPT-3.5) to translate our datasets, followed by evaluations in zero-shot, one-shot, and CoT settings for GPT-3.5. To streamline our conclusions, we report the average performance of these prompting techniques across languages and figurative types in Figure 3.

For idioms, GPT-3.5 outperforms both Google Translate and NLLB for most high-resource or Latin languages (e.g., In, Zh, De, and Fa). However, for similes, Google and NLLB perform better than GPT-3.5 in lowest-resource languages (Jv, Kn, Su). This result aligns with our expectations, as these models were trained to provide better translations for a wider range of languages, including low-resource ones. This finding also mirrors the results of Liu et al. (2024a), which highlight that while NLLB shows strong performance, Google Translate tends to outperform it in most scenarios when handling multilingual tasks. Overall, translation with Google is the better choice for similes, while GPT-3.5 excels with idioms. This distinction may arise from the fact that literal translations by Google are more likely to alter the meaning of idioms compared to similes.

5 Analysis

In this section, we present further analyses of our observations to deepen our comparative study of simile and idiom interpretation.



(a) Simile

(b) Idiom

Figure 3: Comparing the translation methods using GPT-3.5. The average accuracy of zero-shot, one-shot, and CoT methods is reported for each translation method.

5.1 The thought pathway in CoT

We investigate the thought process and pathway of the LLMs doing CoT in our experiments. We analyze instances where models initially generate incorrect answers in a zero-shot setting but provide accurate responses when using CoT. We focus on closed-source models here.

For idiom interpretation, we observe that models predominantly generate responses and explanations in English, even when the one-shot example is presented in a different native language. In contrast, for simile interpretation, the behavior varies across models: GPT-4o mini consistently responds in the native language, while GPT-3.5 produces a mix of English and native language responses, with the ratio of English outputs varying depending on the language. Gemini 1.5, however, tends to generate explanations primarily in English, which can lead to misinterpretations of culturally specific concepts. Detailed results of language detection in CoT responses are provided in Table 16.

To further investigate the impact of language in CoT responses on simile interpretation, we conducted additional experiments. We explicitly prompted Gemini 1.5 to generate responses in the native language, and instructed GPT-4o mini to respond in English (Appendix F). While GPT-4o mini showed no significant performance differ-

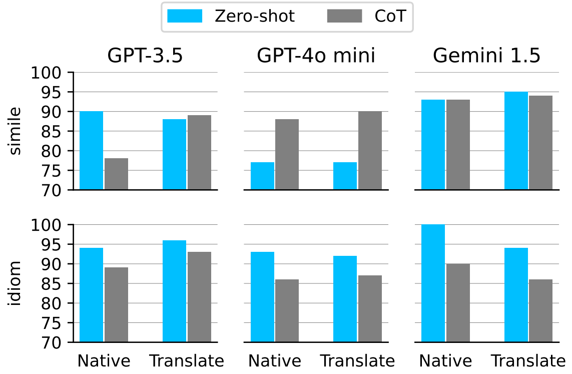


Figure 4: Comparing consistency (%) of closed-source models in zero-shot and CoT settings when input is in native or Translated-English. Each number represents the average on languages.

ences when responding in English (except Sw), Gemini 1.5 displayed improved performance in Fa, Su, and Jv when responding in their respective native languages. Results from these new experiments are presented in Table 17 in the Appendix.

5.2 Dialogue simulation

We evaluate the dialogue simulation technique for simile interpretation, expecting context to improve understanding (Liu et al., 2024b). However, results show that CoT prompting outperforms dialogue simulation across almost all languages (Table 5). This shows that reasoning and then explicitly deriving the meaning of a simile (as in the one-shot CoT setting) is a better approach than using it in a context like a conversation.

lang.	Id	Hi	Sw	Jv	Kn	Su	Fa
CoT	**.864**	**.685**	**.818**	**.735**	**.705**	.678	**.855**
Dial.	.803	.661	.751	.698	.538	**.696**	.806

Table 5: Comparison of GPT-3.5 performance using CoT and dialogue simulation techniques in translated-English prompts. *Dial. refers to dialogue simulation.

5.3 Consistency of results

We examine the reliability of some prompting methods in answering questions with *consistency* metric, i.e. the proportion of samples where the model gives the same answer across all three runs, regardless of whether the answer is correct or not.

We employ the consistency metric to show the uncertainty and reliability of the models. By examining how consistency shifts from zero-shot to CoT in idiom, we observe that models generally exhibit increased uncertainty when they generate an explanation before providing an answer. However, the effect of CoT varies across different models for simile. Specifically, we find that GPT-4o mini demonstrates greater reliability after using CoT, while GPT-3.5 in native prompting shows a decrease. In contrast, Gemini 1.5 maintains a steady level of consistency throughout.

5.4 Making idiom examples more challenging

To better understand the limitations of the datasets better, we conduct an analysis on the Persian idiom dataset. We modify 60 samples by replacing one of the answer choices with the literal meaning of the idiom while retaining the correct non-literal meaning. This approach tests the model's ability to distinguish between figurative and literal interpretations. For instance, as shown in Table 18, the idiom "آب از دست کسی نچکیدن" which literally means "water does not drip from someones hand", is a figurative expression meaning "someone is very stingy". The results indicate an 8% absolute decrease in accuracy for GPT-4o mini (i.e., a drop in 4 out of 60 examples), underscoring the model's challenges in accurately interpreting idiomatic expressions when presented with plausible literal alternatives. This finding emphasizes the inherent complexity of figurative language understanding for LLMs.

6 Conclusion

This study examined how multilingual LLMs interpret similes and idioms across languages, comparing open-source and closed-source models using MABL, MAPS, and newly developed Persian datasets. We tested prompt engineering strategies like one-shot, CoT, and dialogue simulation, finding their effectiveness varied by figurative language type, language, and model. While open-source models like Llama 3.1 and Qwen2 performed well overall, they struggled with similes in low-resource languages. Idiom interpretation, however, showed near-saturation, highlighting the need for more challenging datasets.

Our two new Persian datasets contribute valuable resources for evaluating LLMs in this language. Expanding the scope of figurative language types to include metaphors, sarcasm, and irony could provide a more comprehensive evaluation of LLMs' capabilities. Also, developing datasets that challenge LLMs with more context-dependent or ambiguous figurative expressions will be crucial for driving progress in this area.

7 Limitation

While our study offers valuable insights into multilingual figurative language understanding, several limitations remain. First, we primarily focus on similes and idioms, excluding other important figurative types like metaphors, sarcasm, and irony due to the scarcity of relevant datasets. Additionally, the datasets used in this research cover different languages, with only English and Indonesian being common across both, complicating cross-language comparisons. The dataset quality, especially for low-resource languages, also could not be verified by native speakers, potentially introducing inaccuracies in culturally specific expressions. Moreover, many open-source models, including Llama 3.1 and Qwen2, lack official support for low-resource languages like Sundanese and Javanese, making their performance in these contexts less reliable. Lastly, the datasets may not be challenging enough for advanced models, as GPT-4o mini and Gemini 1.5 Flash nearly achieve perfect accuracy in high-resource languages, pointing to the need for more complex and context-dependent figurative tasks to fully test model capabilities.

Acknowledgements

We extend our gratitude to Yong Zheng-Xin and Emmy Liu for their valuable feedback and insightful comments on an earlier version of this paper. We also sincerely thank the anonymous reviewers for their valuable feedback and thoughtful suggestions.

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6455–6469, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, and Junteng Jia et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Susan R Fussell and Mallie M Moss. 2014. Figurative language in emotional communication. In *Social and cognitive approaches to interpersonal communication*, pages 113–141. Psychology Press.

Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. 2017. The role of conversation context for sarcasm detection in online interactions. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 186–196, Saarbrücken, Germany. Association for Computational Linguistics.

Tomáš Hercig and Ladislav Lenc. 2017. The impact of figurative language on sentiment analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 301–308, Varna, Bulgaria. INCOMA Ltd.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Kaiyu Huang, Fengran Mo, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, et al. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936*.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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.

Huiyuan Lai and Malvina Nissim. 2022. Multi-figurative language generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5939–5954, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9254–9267, Toronto, Canada. Association for Computational Linguistics.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Preprint*, arXiv:2402.10946.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *Preprint*, arXiv:2405.15145.

Linlin Li and Caroline Sporleder. 2009. Classifier combination for contextual idiom detection without labelled data. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 315–323, Singapore. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav

Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Changsheng Liu and Rebecca Hwa. 2016. Phrasal substitution of idiomatic expressions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 363–373, San Diego, California. Association for Computational Linguistics.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024a. Is translation all you need? a study on solving multilingual tasks with large language models. *Preprint*, arXiv:2403.10258.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024b. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. 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 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Lizhen Liu, Xiao Hu, Wei Song, Ruiji Fu, Ting Liu, and Guoping Hu. 2018. Neural multitask learning for simile recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1543–1553, Brussels, Belgium. Association for Computational Linguistics.

Suzanne Mpouli. 2017. Annotating similes in literary texts. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.

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.

Jing Peng and Anna Feldman. 2016. Experiments in idiom recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2752–2761, Osaka, Japan. The COLING 2016 Organizing Committee.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Ashequl Qadir, Ellen Riloff, and Marilyn Walker. 2015. Learning to recognize affective polarity in similes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 190–200, Lisbon, Portugal. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Richard M Roberts and Roger J Kreuz. 1994. Why do people use figurative language? *Psychological science*, 5(3):159–163.

Giancarlo Salton, Robert Ross, and John Kelleher. 2016. Idiom token classification using sentential distributed semantics. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 194–204, Berlin, Germany. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Ekaterina V Shutova. 2011. Computational approaches to figurative language. Technical report, University of Cambridge, Computer Laboratory.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *Preprint*, arXiv:2409.12183.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, and Chih-Kuan Yeh et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff

Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Rakesh Verma and Vasanthi Vuppuluri. 2015. A new approach for idiom identification using meanings and the web. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 681–687, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Shun Wang, Ge Zhang, Han Wu, Tyler Loakman, Wenhao Huang, and Chenghua Lin. 2024a. Mmte: Corpus and metrics for evaluating machine translation quality of metaphorical language. *arXiv preprint arXiv:2406.13698*.

Xiaolong Wang, Yile Wang, Yuanchi Zhang, Fuwen Luo, Peng Li, Maosong Sun, and Yang Liu. 2024b. Reasoning in conversation: Solving subjective tasks through dialogue simulation for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15880–15893, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Zhiwei Yu and Xiaojun Wan. 2019. How to avoid sentences spelling boring? towards a neural approach to unsupervised metaphor generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 861–871, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiali Zeng, Linfeng Song, Jinsong Su, Jun Xie, Wei Song, and Jiebo Luo. 2019. Neural simile recognition with cyclic multitask learning and local attention. *Preprint*, arXiv:1912.09084.

Jiayi Zhang, Zhi Cui, Xiaoqiang Xia, Yalong Guo, Yanran Li, Chen Wei, and Jianwei Cui. 2021. Writing polishment with simile: Task, dataset and a neural approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 16, pages 14383–14392.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Preprint*, arXiv:2306.05179.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics.

A Simile and Idiom Datasets

Here is a detailed explanation of the MABL and MAPS datasets, designed to evaluate figurative language understanding for similes and idioms, respectively.

1.1 MABL for Simile

The MABL dataset is designed to assess figurative language understanding, consists of similes. Similes are rhetorical devices that compare two different concepts, often using words like "as" or "like", to highlight similarities. Understanding similes requires a nuanced grasp of both literal and figurative meanings within a sentence, making this a challenging task for language models.

The primary task in the MABL dataset is a binary classification problem. Each instance, as shown in Table 1, consists of: **start phrase**, a sentence containing a simile, Two possible continuations, **ending 0** and **ending 1**, where one reflects the correct figurative meaning and the other is often opposing or incorrect, and A binary **label** indicating the correct ending. This task evaluates whether language models can correctly interpret the intended figurative meaning of similes.

1.2 MAPS for Idiom

The MAPS dataset is designed to evaluate the understanding and interpretation of proverbs and idioms within conversational contexts. Proverbs and idioms are commonly used expressions that convey figurative meanings, which often differ significantly from their literal interpretations. Accurate comprehension requires models to infer the intended figurative meaning of these expressions based on context.

The MAPS dataset also presents a binary classification problem. Each instance, as illustrated in Table 2, consists of: **proverb**, a commonly used saying or idiomatic expression, **conversation**, a short dialogue where the proverb or idiom is used, two possible interpretations of the given conversation, one aligned with the figurative meaning of the proverb, and the other aligned with a literal or incorrect interpretation, and binary **label** (A or B) indicating the correct interpretation. This task assesses whether language models can accurately interpret the figurative meaning of idioms and proverbs when presented within a conversational context.

B Dataset Construction

In this section, we provide additional details on the construction of the Persian simile and idiom datasets.

2.1 Persian Simile Dataset

To create the Persian Simile dataset, we task GPT-4o with generating over 800 simile pairs, following the structure used in the MABL dataset. Each sample consists of a start phrase with two possible endings: one correctly reflecting the intended meaning and the other presenting an incorrect or opposite interpretation. Table 6 provides one example of used prompts.

Sample Refinement. After generation, two native Persian speakers review all the samples, focusing on grammatical accuracy, cultural relevance, and the quality of the figurative language. From the 800+ generated samples, the reviewers select approximately 200 of the most relevant and high-quality examples. They further analyze the selected samples, refining those that require adjustments. The refinement process includes correcting figurative meanings and incorporating cultural references and specific characters to enhance the dataset's complexity and relevance for Persian speakers. The reviewers validate both the correctness of the simile usage and the plausibility of the incorrect endings, ensuring that they are clearly distinguishable yet realistic.

The final Persian Simile dataset contains 200 samples, each validated for correctness and cultural appropriateness, making it a robust resource for evaluating simile comprehension in Persian.

Your task is to generate pairs of sentences with opposite or very different meanings, both of which contain Persian similes. You can feel free to incorporate creativity into the similes, but also make sure that they're something that could be understood by the speakers of the language that you are generating similes for, e.g., "this is as classic as pancakes for breakfast" to mean "this is classic" wouldn't make sense for a culture in which pancakes aren't traditionally eaten for breakfast. You can do this by thinking of a simile that conveys a certain meaning, and replacing the similative phrase with another similative phrase of the same type that conveys the opposite meaning. Here are some examples of similes to give you an idea of what we're looking for: Please write the simile and two correct and incorrect meanings. ۱. دست‌هایش مثل تنه‌ی درخت بلوط، قوی و محکم هستند. (Her hands are as strong and sturdy as the trunk of an oak tree.) دست‌هایش ضعیف و بی‌رمق هستند / دست‌هایش قوی و پرتوان هستند (Her hands are weak and frail / Her hands are strong and powerful) ۲. رفتارش مثل نسیمی است که در میان گل‌ها می‌وزد (Her behavior is like a breeze blowing among the flowers) رفتارش پرخاشگرانه و تند است./رفتارش آرام و متین است. (Her behavior is aggressive and abrupt / His behavior is calm and composed)

Table 6: An example of the prompt used in the construction of the Persian simile dataset.

2.2 Persian Idiom Dataset

To create the Persian Idiom dataset, we first collect over 400 idioms along with their meanings from two online resources, Daneshchi and Abadis. From these, we remove any idioms deemed inappropriate or unsuitable for our purposes. We then use GPT-4o to generate conversational contexts for the remaining idioms, ensuring that the idioms are correctly integrated within natural dialogues.

Sample Refinement. Two native Persian speakers review the generated conversational contexts, refining them to ensure grammatical accuracy, cultural relevance, and correct idiomatic usage. During this process, we incorporate cultural references, including historical characters, cultural concepts, and significant events—such as the great wars in the history of Iran—to increase the dataset's complexity and make it more representative of Persian culture, similar to the approach used in the MABL dataset.

Option Generation. In the second round, GPT-4o generates two response options for each conversational context: one correct and one incorrect, based on the meaning of the idiom within the dialogue. Once again, two native Persian speakers review the generated responses, ensuring that the grammar, idiom usage, and cultural aspects are accurate and appropriate.

The final Persian Idiom dataset contains 316 samples, each with a conversational context and two response options. It serves as a valuable resource for evaluating idiomatic comprehension in Persian, emphasizing cultural and linguistic accuracy.

C Cost For Running Experiments

We access these models through the APIs provided by OpenAI [4], Gemini [5], and OpenRouter [6] for accessing open-source models. The total cost for running the experiments is estimated to be under $40, with approximately $30 spent on OpenAI and $10 on OpenRouter.

D Prompt Templates

In this section, we provide the prompting techniques templates used for different tasks, such as zero-shot, one-shot, and dialogue simulation (RiC).

4.1 Zero-shot

This template is designed to assess the model's basic understanding without providing any examples. The instructions are in English, while the input may be in the native language or English. The prompt templates for MABL and MAPS samples are shown in Tables 7 and 8.

In this task, you are given a start phrase indicating a figurative expression in <language> culture. Please select 0 if the start phrase conveys the meaning of ending 0, and 1 if it conveys the meaning of ending 1.

Start Phrase: <start phrase>
Ending 0: <ending 0>
Ending 1: <ending 1>
Answer:

Table 7: Zero-shot prompt template for MABL samples.

Question: How would one interpret this proverb in <language> culture, given the context? Please choose between A and B.
Proverb: <proverb>
Context: <context>
Choices: A: <answer A> B: <answer B>
Answer:

Table 8: Zero-shot prompt template for MAPS samples.

4.2 One-shot

In the one-shot template, a single example is provided to guide the model in understanding the task. The instruction is given in English, while the input language can vary between native and translated-English (as shown in Tables 9 and 10), or a combination of both native and translated-English (Tables 11 and 12).

In this task, you are given a start phrase indicating a figurative expression in <language> culture. Please select 0 if the start phrase conveys the meaning of ending 0, and 1 if it conveys the meaning of ending 1.

Below is an example showing you how to do the task:
Start Phrase: <sample start phrase>
Ending 0: <sample ending 0>
Ending 1: <sample ending 1>
Answer: <0/1>

Now answer the following question:
Start Phrase: <start phrase>
Ending 0: <ending 0>
Ending 1: <ending 1>
Answer:

Table 9: One-shot prompt template for MABL samples (Native or Translated-English).

4.3 Chain of Thought (CoT)

This template is designed to test the model's ability to reason through the task step-by-step before arriving at a final decision. The prompt includes an example that demonstrates how the model should generate both an answer and the reasoning behind it. The CoT prompt structure for similes and idioms are detailed in Table 13 and Table 14 respectively.

4.4 Dialogue Simulation (RiC)

This template is used in the RiC method, where the model generates a conversation that includes the figurative expression, helping it understand the phrase in context. The prompt template for MABL samples is shown in Table 15.

[4] https://platform.openai.com
[5] https://ai.google.dev
[6] https://openrouter.ai

Question: How would one interpret this proverb in <language> culture, given the context? Please choose between A and B.
Proverb: <sample proverb>
Context: <sample context>
Choices: A: <sample answer A> B: <sample answer B>
Answer: <A/B>

Question: How would one interpret this proverb in <language> culture, given the context? Please choose between A and B.
Proverb: <proverb>
Context: <context>
Choices: A: <answer A> B: <answer B>
Answer:

Table 10: One-shot prompt template for MAPS samples (Native or Translated-English).

In this task, you are given a start phrase indicating a figurative expression in <language> culture. Please first translate the start phrase, ending 0, and ending 1 into English. Then, select 0 if the translated start phrase conveys the meaning of the translated Ending 0, and 1 if it conveys the meaning of the translated Ending 1.

Below is an example showing you how to do the task:
Start Phrase: <sample start phrase>
Ending 0: <sample ending 0>
Ending 1: <sample ending 1>

Translated into English:
Start Phrase: <sample start phrase English translation>
Ending 0: <sample ending 0 English translation>
Ending 1: <sample ending 1 English translation>
Answer: <0/1>

Now answer the following question:
Start Phrase: <start phrase>
Ending 0: <ending 0>
Ending 1: <ending 1>
Translated into English:
Answer:

Table 11: One-shot prompt template for MABL samples (Native and Translated-English).

Question: How would one interpret this proverb in <language> culture, given the context? Please first translate the Proverb, Context, and Choices into English. Then, choose between A and B.
Proverb: <sample proverb>
Context: <sample context>
Choices: A: <sample answer A> B: <sample answer B>

The English translation is:
Proverb: <sample proverb English translation>
Context: <sample context English translation>
Choices: A: <sample answer A English translation> B: <sample answer B English translation>
Final Answer: <A/B>

Question: How would one interpret this proverb in <language> culture, given the context? Please first translate the Proverb, Context, and Choices into English. Then, choose between A and B.
Proverb: <proverb>
Context: <context>
Choices: A: <answer A> B: <answer B>

The English translation is:
Final Answer:

Table 12: One-shot prompt template for MAPS samples (Native and Translated-English).

E Language detection of CoT responses of closed-source models

We use the Google Translate Python library for language detection of CoT responses in the native prompting setting for closed-source models in both simile and idiom. Results are in Table 16.

F Examine the influence of CoT responses language

As the Gemini 1.5 and GPT-4o mini exhibit different behaviors in the language used for generating re-

In this task, you are given a start phrase indicating a figurative expression in <language> culture. Please select 0 if the start phrase conveys the meaning of ending 0, and 1 if it conveys the meaning of ending 1.

Below is an example showing you how to do the task:
Start Phrase: <sample start phrase>
Ending 0: <sample ending 0>
Ending 1: <sample ending 1>
Answer: <sample answer with reasoning>

Now answer the following question:
Start Phrase: <start phrase>
Ending 0: <ending 0>
Ending 1: <ending 1>
Answer:

Table 13: CoT prompt template for MABL samples.

Question: How would one interpret this proverb in <language> culture, given the context? Please first think about the proverb's meaning, then write an explanation of the proverb's meaning, and finally choose between A and B.
Proverb: <sample proverb>
Context: <sample context>
Choices: A: <sample answer A> B: <sample answer B>
Explanation: <sample explanation>
Answer: <A/B>

Question: How would one interpret this proverb in <language> culture, given the context? Please first think about the proverb's meaning, then write an explanation of the proverb's meaning, and finally choose between A and B.
Proverb: <proverb>
Context: <context>
Choices: A: <answer A> B: <answer B>
Explanation:
Answer:

Table 14: CoT prompt template for MAPS samples.

Figurative Language Interpretation: In this task, you are given a start phrase indicating a figurative expression in <language> culture. Please select 0 if the start phrase conveys the meaning of ending 0, and 1 if it conveys the meaning of ending 1.

Start Phrase: <start phrase>
Ending 0: <ending 0>
Ending 1: <ending 1>

First, extract keywords from the question.
Then, according to the keywords, construct a scenario for the question in the form of dialogue.
Finally, according to the question and conversation, reason and give the final answer. Select from 0 or 1.

Table 15: Dialogue simulation (RiC) prompt template for MABL samples.

sponses for CoT, two additional experiments were conducted to examine the influence of language used in the responses. In the new experiments, the phrase "*Give the Answer in <language> language*" was appended to the end of the prompts. For Gemini, the model was asked to respond in its native language of example, while GPT-4o was instructed to respond in English. The accuracy results for the new prompts are presented in Table 17.

G Distinguish Between Figurative and Literal Meaning of Idioms

Here is an example of the prompt structure in Table 18.

	S	GPT-4o mini	GPT3.5	Gemini 1.5
		Simile		
En	200	100	100	100
Hi	200	0	3.5	91.4
Id	200	0	10	45.5
Sw	200	0	4.5	98.9
Jv	200	5.5	46.5	93.2
Kn	200	2	35.7	92.1
Su	200	0	0.5	88.9
Fa	200	0	13	51.4
		Idiom		
En	206	100	100	100
Zh	139	96.5	95.1	100
Bn	262	93.7	24.7	95
Ru	213	99.6	65.5	98.6
De	172	95.1	99.5	100
Id	248	93.6	45.3	99.6
Fa	299	74.1	13	88.3

Table 16: Ratio (%) of CoT examples in which their languages are detected as English. Examples of native prompting in the CoT setting are considered. S refers to the number of samples of each language.

Lang.	Gemini 1.5		GPT-4o mini	
	Default (English)	New (Native)	Default (Native)	New (English)
En	.887	.900	.916	.925
Id	.924	.910	.911	.913
Hi	.744	.695	.685	.696
Sw	.804	.775	.768	.803
Jv	.670	.760	.783	.794
Kn	.622	.585	.576	.580
Su	.745	.790	.768	.766
Fa	.825	.925	.898	.896
Avg	.777	.792	.788	.797

Table 17: Performance (%) of Gemini 1.5 when prompted to respond in native, and GPT-4o minie when prompted to respond in English. Experiments are on the simile dataset for native prompting and CoT setting.

Question: What is the meaning of the following phrase in Persian, given the context? Please choose between A and B.

Phrase: آب از دست کسی نچکیدن

(Water does not drip from someone's hand)

Context: شخص ۱: از اون خواستم مقداری پول به من قرض دهد. شخص ۲: اون را می‌شناسم. آب از دستش نمی‌چکد!

(Person 1: I asked them to lend me some money. Person 2: I know them. Water does not drip from their hand!)

Choices:

A: فردی خسیس و تنگ نظر بودن

(Being stingy and narrow-minded)

B: محافظ خوبی برای آب بودن

(Being a good guardian of water)

Answer:

Table 18: Prompt structure used for analyzing the model's ability to distinguish between figurative and literal meanings. Choice A indicates figurative meaning, while Choice B provides a plausible literal interpretation. Note: English translations, provided in parentheses below the original Persian phrases, are not part of the prompt presented to the model.