MasalBench: A Benchmark for Contextual and Cross-Cultural Understanding of Persian Proverbs in LLMs

Ghazal Kalhor[1] **Behnam Bahrak**[2]
[1]EPFL [2]Khatam University
Correspondence: ghazal.kalhor@epfl.ch, b.bahrak@teias.institute

Abstract

In recent years, multilingual Large Language Models (LLMs) have become an inseparable part of daily life, making it crucial for them to master the rules of conversational language in order to communicate effectively with users. While previous work has evaluated LLMs' understanding of figurative language in high-resource languages, their performance in low-resource languages remains underexplored. In this paper, we introduce MASALBENCH, a comprehensive benchmark for assessing LLMs' contextual and cross-cultural understanding of Persian proverbs, which are a key component of conversation in this low-resource language. We evaluate eight state-of-the-art LLMs on MASALBENCH and find that they perform well in identifying Persian proverbs in context, achieving accuracies above 0.90. However, their performance drops considerably when tasked with identifying equivalent English proverbs, with the best model achieving 0.79 accuracy. Our findings highlight the limitations of current LLMs in cultural knowledge and analogical reasoning, and they provide a framework for assessing cross-cultural understanding in other low-resource languages. MASALBENCH is available at https://github.com/kalhorghazal/MasalBench.

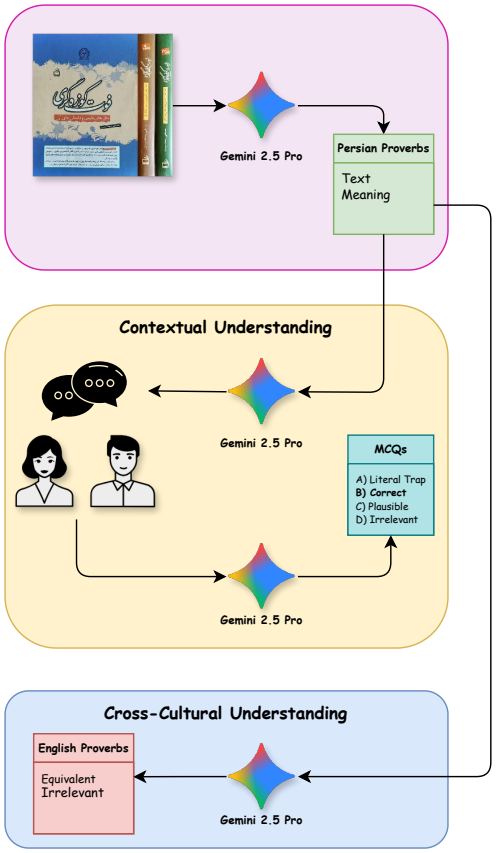


Figure 1: Pipeline illustrating the construction of the benchmark for assessing LLMs' contextual and cross-cultural understanding of Persian proverbs.

1 Introduction

Large Language Models (LLMs) are increasingly used in daily life, from offering product advice to answering general questions and supporting business tasks. To communicate effectively, they must understand users' conversational language. Proverbs, traditional sayings that convey commonsense truths metaphorically, are an important aspect of natural conversation. Multilingual LLMs perform well on proverbs in high-resource languages, particularly English, thanks to abundant training data and their inclusion in benchmarks (Kim et al., 2025; Fu et al., 2025; Almeida et al., 2025; Liu et al., 2024). In contrast, the scarcity of such data in low-resource languages may limit LLM proficiency with their proverbs, and only a few studies develop benchmarks to investigate this capability, typically focusing on a small and limited set of low-resource languages (Azime et al., 2025; Thapa et al., 2025; Donthi et al., 2025; Nguyen and Kordjamshidi, 2025; Das et al., 2026).

Persian is one of the low-resource languages for multilingual LLMs, yet it is exceptionally rich in the use of proverbs, both historically and in contemporary contexts. Iranian speakers naturally employ

Proceedings of the First Workshop on Pluralistic Value Alignment of LLMs, pages 17–23
November 10, 2026 ©2026 Association for Computational Linguistics

proverbs frequently in daily conversations and even in professional settings, often without being consciously aware of it (Hadissi, 2010). However, only a few studies examine LLMs' ability to understand Persian proverbs. For example, Khoshtab et al. (2024) propose 316 binary-choice questions in dialogue format, where each question pairs the correct interpretation of a proverb with a single incorrect option generated by GPT-4o. In 60 of these questions, the incorrect option corresponds to the literal meaning of the proverb, making the evaluation of contextual and figurative understanding more challenging. Similarly, Farsi et al. (2025) design 370 multiple-choice questions that ask for the meaning of Persian proverbs, where the distractors are meanings of randomly selected proverbs.

To address this gap, we introduce MASAL-BENCH, a comprehensive benchmark for evaluating LLMs' understanding of Persian proverbs at both contextual and cross-cultural levels. At the contextual level, we propose a set of 1,000 multiple-choice questions in dialogue form, where each of the three distractor options is carefully designed to serve a specific purpose: a literal trap, a plausible but incorrect meaning, and an irrelevant meaning. At the cross-cultural level, we propose 700 binary-choice questions to assess whether LLMs can identify the English equivalent of a given Persian proverb, with the distractor option crafted to share similar wording, tone, or imagery but convey a different meaning. While previous work explore proverb translation (Donthi et al., 2025; Rezaeimanesh et al., 2025), to the best of our knowledge, no prior study has captured the cross-cultural dimension through systematic equivalence pairing between languages. Finally, we evaluate MASAL-BENCH on eight prominent multilingual LLMs.

Building on these contributions, this study investigates three research questions: **RQ1:** How well do multilingual LLMs understand Persian proverbs in context? **RQ2:** What kinds of mistakes do they make when failing to do so, and how often do these correspond to literal, plausible, or irrelevant errors? **RQ3:** To what extent can LLMs identify the correct English equivalent of a given Persian proverb?

Our results indicate that multilingual LLMs perform strongly in understanding Persian proverbs in context, with model size and instruction tuning being key contributors. When errors occur, they predominantly involve plausible distractors, reflecting the models' tendency to select semantically coherent interpretations, a pattern more pronounced in reasoning-focused models. In contrast, identifying the correct English equivalent of a Persian proverb proves more challenging, and although instruction tuning provides some benefit, its impact is limited due to the need for cultural knowledge and analogical reasoning.

2 Benchmark Construction

To construct our benchmark, we collect Persian proverbs and their meanings from *Foote Koozegari: Persian Proverbs and Their Stories* (Rahmandoost et al., 2011), a comprehensive reference that documents the etymologies, narratives, and interpretations of over 4,000 Persian proverbs. For this task, we use Gemini 2.5 Pro (Comanici et al., 2025) with the PDF versions of the book (both volumes), as the texts are provided in image format rather than as digital text.[1] We choose Gemini because it has strong performance on Persian text recognition and understanding,[2] making it well suited for processing scanned Persian texts. To ensure consistency during benchmark construction, we use a single model and manually review the extracted content to correct potential typos. We then select 1,000 proverbs that are widely used and well known in Iranian culture, based on the authors' cultural familiarity and their frequency of occurrence in everyday language, literature, and media. All manual reviews and verifications are carried out by the authors, who are native Persian speakers familiar with the language and its cultural context. The overall pipeline for constructing the benchmark is illustrated in Figure 1.

2.1 Contextual Understanding

To prepare the dataset for this task, we follow two steps. First, for 100 of our proverbs, we manually design short, natural, and conversational dialogues in two parts between two people. In each dialogue, the second person uses the given proverb in response to the first person. We also provide an explanation of the second person's purpose for using the proverb. We then automate this step for the remaining proverbs using Gemini 2.5 Pro, by providing the proverbs and their meanings. Second, for 40 dialogues, we manually design three distractor interpretations of the proverb: (1) a literal

[1] Whenever we use Gemini 2.5 Pro for automated content generation, we set the temperature to 1 and top-p to 0.95 to ensure a balance of creativity and coherence.

[2] https://mcinext-mizan-llm-leaderboard.hf.space/

interpretation, (2) a logical but incorrect interpretation, and (3) an unrelated interpretation. We then automate this step for the remaining dialogues using Gemini, by providing the dialogue, the question (asking for the second person's purpose for using the proverb), and the correct purpose. All automatically generated dialogues and distractor options are manually verified by the authors for correctness and naturalness. These steps result in a dataset of 1,000 multiple-choice questions for assessing LLMs' understanding of proverbs in context.

2.2 Cross-Cultural Understanding

To construct the dataset for this task, we provide Gemini 2.5 Pro with each Persian proverb and its meaning, and instruct it to generate two well-known and recognized English proverbs: (1) an equivalent proverb and (2) an irrelevant proverb. The equivalent proverb is most similar in meaning and usage to the given Persian proverb, while the irrelevant proverb has no real connection in meaning or usage to the Persian proverb, but may appear similar in form (for example, in terms of words, tone, or imagery), and therefore serves as a convincing distractor. If no clear or culturally appropriate English equivalent exists for a given Persian proverb, we exclude that proverb from this task and instead sample additional proverbs from the source collection. We then manually verify the results to ensure that the generated proverbs are authentic English expressions and that the equivalents accurately match the Persian proverbs in meaning. Additionally, for each proverb, we add a question asking which English proverb is equivalent to it. Through this process, we compile a dataset of 700 binary-choice questions that assesses LLMs' cross-cultural understanding of proverbs.

3 Experiments

3.1 Language Model for Evaluation

We evaluate several well-known multilingual LLMs of varying sizes with proficiency in Persian, including Llama 4 Scout (AI, 2025), Llama 3.3 70B Instruct (Meta AI, 2024), Qwen 2.5 72B Instruct (Qwen Team, 2025a), Qwen QwQ 32B (Qwen Team, 2025b), DeepSeek V3.1 (DeepSeek, 2025), DeepSeek R1 (DeepSeek-AI, 2025), GPT-4.1 mini (OpenAI, 2025), and GPT-4o mini (OpenAI, 2024), on our benchmark.

3.2 Prompting Setup

To design the prompts for both tasks, we randomize the order of the options to avoid positional bias (Pezeshkpour and Hruschka, 2024). We employ zero-shot prompting, and each prompt runs independently, with the LLMs not performing any web searches to answer them. We set the temperature to 0 to ensure deterministic responses without added creativity, and top-p to 1 so that all possibilities are considered. The maximum number of tokens is limited to 5, as the LLMs are expected to return only the letter corresponding to the correct option. Appendix A provides example prompts used to construct MASALBENCH and evaluate the performance of LLMs on this benchmark.

3.3 Evaluation Metrics

We use accuracy as the primary evaluation metric. Since both tasks are single-label classification problems in which each example has exactly one gold label and each model prediction corresponds to a single label, micro-F1 is mathematically equivalent to accuracy. Therefore, we report accuracy only.

4 Main Results

4.1 Contextual Level

Table 1 presents the accuracy of different LLMs on contextual understanding. As shown, all models perform highly on this task, achieving accuracy above 0.90, which indicates that contextual understanding of Persian proverbs is a relatively easy task for these state-of-the-art LLMs and that they are already robust at metaphorical reasoning in non-English contexts. DeepSeek V3.1 turns out to be the best-performing model with an accuracy score of 0.943, which could be explained by the fact that it has the largest size (671B parameters) among the models. However, size is not the only determining factor of LLM performance on this task, as the lowest accuracy score is assigned to Qwen QwQ 32B, even though it has a larger parameter size compared to the GPT models. We also observe that instruction-tuned models perform better than their reasoning-focused counterparts within the same families, which suggests that instruction tuning plays a significant role in improving cultural contextualization.

To assess the types of mistakes LLMs make when failing to select the correct meaning of a Persian proverb in context, we analyze the proportion of responses corresponding to each distractor

Model	Accuracy
Llama 4 Scout	0.906
Llama 3.3 70B Instruct	0.927
Qwen 2.5 72B Instruct	0.911
Qwen QwQ 32B	0.903
DeepSeek V3.1	**0.943**
DeepSeek R1	0.927
GPT-4.1 mini	0.920
GPT-4o mini	0.920

Table 1: Accuracy values for contextual understanding across LLMs.

option. As shown in Figure 2, the majority of errors across all models fall into plausible distractors. This suggests that when LLMs are uncertain about a proverb's meaning, they still attempt to choose a semantically coherent interpretation relevant to the dialogue. We also observe that within the same model family, reasoning-focused models are more likely to fall for plausible distractors compared to instruction-tuned models. In contrast, literal and irrelevant distractors occur at much lower rates (ranging from 0.02 to 0.00), indicating that these advanced LLMs rarely make such simplistic mistakes.

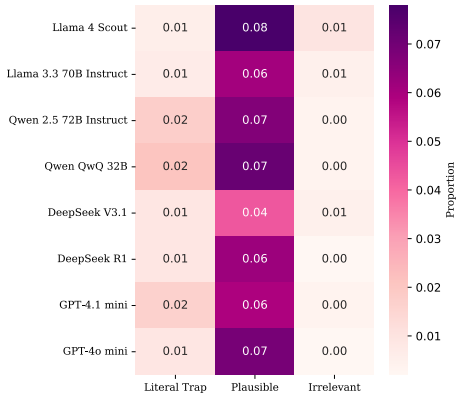


Figure 2: Heatmap of the proportion of distractor option selections for contextual understanding across LLMs.

4.2 Cross-Cultural Level

Table 2 presents the accuracy of different LLMs on cross-cultural understanding. As shown, LLMs perform worse on this task compared to contextual understanding. This may be partly explained by the fact that cross-cultural understanding requires cultural knowledge beyond linguistic decoding and also demands analogical reasoning, since equivalent proverbs often employ different imagery or wording. The best-performing model is DeepSeek R1, with an accuracy score of 0.793, while the lowest-performing model is Llama 4 Scout (0.647). We also observe that instruction tuning continues to provide advantages in this task, although to a lesser extent than in contextual understanding, as success here depends more on training exposure to cultural data than on improved instruction following. Overall, these results suggest that LLMs are relatively strong at intra-language contextual reasoning but weaker at cross-cultural abstraction.

Model	Accuracy
Llama 4 Scout	0.647
Llama 3.3 70B Instruct	0.710
Qwen 2.5 72B Instruct	0.691
Qwen QwQ 32B	0.691
DeepSeek V3.1	0.703
DeepSeek R1	**0.793**
GPT-4.1 mini	0.703
GPT-4o mini	0.651

Table 2: Accuracy values for cross-cultural understanding across LLMs.

5 Conclusion

This work introduces MASALBENCH, a benchmark designed to probe the contextual and cross-cultural understanding of Persian proverbs by LLMs. The dataset includes 1,000 multiple-choice questions targeting speaker intent in context and 700 binary-choice items on matching Persian proverbs to their English equivalents. Our evaluation across eight widely used multilingual LLMs shows that they generally perform well in capturing the intended meaning of Persian proverbs in context, with larger model size and instruction tuning playing a crucial role in this strength. When errors do arise, they are typically due to plausible but incorrect alternatives, highlighting a consistent preference for semantically coherent interpretations, a tendency especially evident in reasoning-oriented models. By contrast, transferring this knowledge across cultures proves substantially harder: aligning Persian proverbs with their English counterparts remains a notable challenge, as it calls for cultural grounding and analogical reasoning beyond what instruction tuning alone can provide. Taken together, MASALBENCH and our findings shed light on current limitations and offer a foundation for advancing research on cross-lingual and culturally informed language understanding.

Limitations

Our study has several limitations that could be addressed in future work. First, we rely solely on a single LLM (Gemini 2.5 Pro) for constructing our benchmark, even though other prominent LLMs with proficiency in Persian exist. Future research could explore alternative LLMs and compare the quality of their generated questions with ours to identify the most suitable model for this task. Second, equivalent English proverbs for Persian proverbs are relatively rare, making our cross-cultural understanding task inherently more challenging for LLMs than the core task itself. Future work could investigate alternative methods for evaluating LLMs' cross-cultural understanding of Persian proverbs.

References

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Meta AI blog (LLaMA 4 announcement).

Thales Sales Almeida, Giovana Kerche Bonás, and João Guilherme Alves Santos. 2025. Broverbs–measuring how much llms understand portuguese proverbs. *arXiv preprint arXiv:2509.08960*.

Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Yonas Chanie, Bontu Fufa Balcha, Negasi Haile Abadi, Henok Biadglign Ademtew, Mulubrhan Abebe Nerea, Debela Desalegn Yadeta, Derartu Dagne Geremew, et al. 2025. Proverbeval: Exploring llm evaluation challenges for low-resource language understanding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6250–6266.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Sarmistha Das, Shreyas Guha, Suvrayan Bandyopadhyay, Salisa Phosit, Kitsuchart Pasupa, and Sriparna Saha. 2026. When meaning isn't literal: Exploring idiomatic meaning across languages and modalities. *arXiv preprint arXiv:2604.10787*.

DeepSeek. 2025. Deepseek-v3.1 release notes. https://api-docs.deepseek.com/news/news250821. Official DeepSeek release / announcement (DeepSeek website / API docs).

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Sundesh Donthi, Maximilian Spencer, Om Patel, Joon Yong Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving llm abilities in idiomatic translation. In *Future of Information and Communication Conference*, pages 361–375. Springer.

Farhan Farsi, Farnaz Aghababaloo, Shahriar Shariati Motlagh, Parsa Ghofrani, MohammadAli Sadraei-Javaheri, Shayan Bali, Amirhossein Shabani, Farbod Bijary, Ghazal Zamaninejad, AmirMohammad Salehoof, et al. 2025. Melac: Massive evaluation of large language models with alignment of culture in persian language. *arXiv preprint arXiv:2508.00673*.

Yicheng Fu, Zhemin Huang, Liuxin Yang, Yumeng Lu, and Zhongdongming Dai. 2025. Chengyu-bench: Benchmarking large language models for chinese idiom understanding and use. *arXiv preprint arXiv:2506.18105*.

Korosh Hadissi. 2010. A socio-historical approach to poetic origins of persian proverbs. *Iranian Studies*, 43(5):599–605.

Paria Khoshtab, Danial Namazifard, Mostafa Masoudi, Ali Akhgary, Samin Mahdizadeh Sani, and Yadollah Yaghoobzadeh. 2024. Comparative study of multilingual idioms and similes in large language models. *arXiv preprint arXiv:2410.16461*.

Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taeuk Kim. 2025. Memorization or reasoning? exploring the idiom understanding of llms. *arXiv preprint arXiv:2505.16216*.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. 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.

Meta AI. 2024. Llama 3.3 70b instruct — official announcement / model card. https://ai.meta.com/blog/future-of-ai-built-with-llama/. Official Meta AI announcement (see model card for LLaMA-3.3-70B-Instruct on Hugging Face).

Luan Thanh Nguyen and Parisa Kordjamshidi. 2025. Don't take it literally! idiom-aware vietnamese translation via in-context learning. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1795–1814.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. OpenAI blog post.

OpenAI. 2025. Introducing gpt-4.1 in the api. OpenAI blog post.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017.

Qwen Team. 2025a. Qwen2.5 technical report. Technical report, Alibaba/Qwen. ArXiv preprint arXiv:2412.15115.

Qwen Team. 2025b. Qwq-32b: Embracing the power of reinforcement learning. https://qwenlm.github.io/blog/qwq-32b/. Qwen Blog (Mar 6, 2025).

Mostafa Rahmandoost, Azar Rezaei, Marjan Keshavarzi-Azad, and Matin Pedrami. 2011. *Foote Koozegari: Persian Proverbs and Their Stories*. Entesharat e Madreseh, Tehran, Iran.

Sara Rezaeimanesh, Faezeh Hosseini, and Yadollah Yaghoobzadeh. 2025. Large language models for persian-english idiom translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7974–7985.

Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Surabhi Adhikari, Imran Razzak, and Usman Naseem. 2025. Probing the limits of multilingual language understanding: Low-resource language proverbs as llm benchmark for ai wisdom. In *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025)*, pages 120–129.

A Example Prompts

A.1 Prompts for Benchmark Construction

Example prompts used in constructing our benchmark are provided in Figures 3–5, illustrating (1) dialogue generation with Persian proverbs, (2) distractor option generation for multiple-choice questions, and (3) equivalent English proverb generation.

A.2 Prompts for Evaluation

Example prompts used to evaluate our benchmark on LLMs are provided in Figures 6 and 7, illustrating: (1) a multiple-choice question for contextual understanding, and (2) a binary-choice question for cross-cultural understanding.

شما یک مدل زبانی هستید که برای ساخت مکالمه با کاربرد ضرب‌المثل‌های فارسی طراحی شده‌اید.
به شما یک ضرب‌المثل فارسی و کاربرد آن داده می‌شود.
وظیفه شما این است که یک مکالمه‌ی کوتاه، طبیعی و محاوره‌ای در دو بخش بین دو نفر بنویسید. در این مکالمه، نفر دوم باید در پاسخ به نفر اول از ضرب‌المثل داده‌شده استفاده کند. نام شخصیت‌ها باید ایرانی باشند. همچنین لازم است هدف نفر دوم از به‌کار بردن ضرب‌المثل را توضیح دهید.
خروجی را دقیقاً در قالب زیر ارائه کنید:

```
{
  "first_person": "جواد: امیر واقعا دزدی کرده بود؟",
  "second_person": "سیروان: آش نخورده و دهن سوخته.",
  "question": "سیروان با این ضرب‌المثل چه منظوری داشت؟",
  "answer": "امیر دزدی نکرده و بی دلیل متهمش کردن."
}
```

<proverb>شب شراب نیرزد به بامداد خمار.</proverb>
<usage>بسیاری از چیزهایی که در ابتدا خوب و لذت‌بخش و وسوسه‌انگیز به نظر می‌رسند، درنهایت پایانی ناخوش و تلخ دارند.</usage>

English Translation

You are a language model designed to generate conversations that make use of Persian proverbs.
You are given a Persian proverb along with its usage description.
Your task is to write a short, natural, and conversational dialogue in two turns between two speakers. In this dialogue, the second speaker must use the given proverb in response to the first speaker. The characters' names should be Iranian. Additionally, you must explain the intention behind the second speaker's use of the proverb.
Provide the output exactly in the following format:

```
{
  "first_person": "Javad: Did Amir really steal something?",
  "second_person": "Sirvan: He hasn't eaten the stew, yet his mouth is burned.",
  "question": "What did Sirvan mean by using this proverb?",
  "answer": "Amir didn't steal anything and was accused without any reason."
}
```

<proverb>A night of wine is not worth the morning hangover.</proverb>
<usage>Many things that initially appear pleasant, enjoyable, and tempting ultimately lead to an unpleasant and bitter outcome.</usage>

Figure 3: Example prompt for generating dialogues for Persian proverbs, along with its English translation.

شما یک مدل زبانی هستید که برای طراحی گزینه‌های نادرست برای کاربرد ضرب‌المثل‌های فارسی در مکالمه طراحی شده‌اید.
به شما یک مکالمه داده می‌شود که در آن نفر دوم از یک ضرب‌المثل فارسی استفاده کرده است. معنای درست ضرب‌المثل در همان مکالمه نیز داده می‌شود.
وظیفه شما: ساختن سه گزینه نادرست کوتاه، طبیعی و محاوره‌ای:
1. برداشت تحت‌اللفظی از ضرب‌المثل
2. برداشت منطقی اما اشتباه از ضرب‌المثل
3. برداشت غیرمرتبط با مکالمه و ضرب‌المثل
فرمت خروجی:

```
{
  "incorrect_literal": "گزینه تحت‌اللفظی محاوره‌ای",
  "incorrect_plausible": "گزینه منطقی اما اشتباه محاوره‌ای",
  "incorrect_irrelevant": "گزینه غیرمرتبط محاوره‌ای"
}
```

<dialogue> کیان: پریسا خیلی ناراحت شدم شنیدم مهران کل زحمتای پروژه رو به اسم خودش زده. واقعاً ازش انتظار نداشتم.
پریسا: عیبی نداره کیان. گذر پوست به دباغ‌خونه می‌افته.</dialogue>
<question>پریسا با این ضرب‌المثل چه منظوری داشت؟</question>
<correct meaning>بالاخره یه روزی کار مهران به کیان گیر می‌کنه و اون وقته که می‌فهمه چه اشتباهی کرده.</correct meaning>

English Translation

You are a language model designed to generate incorrect options for the task of applying Persian proverbs in conversation.
You are given a dialogue in which the second speaker uses a Persian proverb. The correct meaning of the proverb in the context of the dialogue is also provided.
Your task: Generate three short, natural, and colloquial incorrect options:
1. A literal interpretation of the proverb
2. A logically plausible but incorrect interpretation of the proverb
3. An interpretation that is irrelevant to both the dialogue and the proverb
Output format:

```
{
  "incorrect_literal": "colloquial literal option",
  "incorrect_plausible": "colloquial plausible but incorrect option",
  "incorrect_irrelevant": "colloquial irrelevant option"
}
```

<dialogue>Kian: I was really upset when I heard that Mehran took all the credit for the project and put it under his own name. I really didn't expect that from him.
Parisa: It's okay, Kian. Eventually, the skin passes through the tannery.</dialogue>
<question>What did Parisa mean by using this proverb?</question>
<correct meaning>One day, Mehran will eventually need Kian, and that will be when he realizes the mistake he made.</correct meaning>

Figure 4: Example prompt for generating distractor options in multiple-choice questions for contextual understanding, along with its English translation.

شما یک مدل زبانی هستید که برای پیشنهاد ضرب‌المثل‌های انگلیسی معادل ضرب‌المثل‌های فارسی طراحی شده‌اید.
به شما یک ضرب‌المثل فارسی و کاربرد آن داده می‌شود.
وظیفه شما این است که دو ضرب‌المثل انگلیسی واقعی ارائه دهید:
1. ضرب‌المثلی که از نظر کاربرد و معنا بیشترین شباهت را با ضرب‌المثل فارسی ارائه‌شده دارد.
2. ضرب‌المثلی که از نظر معنا و کاربرد هیچ ارتباطی با ضرب‌المثل فارسی ندارد، اما ممکن است در ظاهر (مثلاً از نظر واژه‌ها، لحن یا تصویرسازی) شباهت‌هایی داشته باشد و به همین دلیل می‌تواند به‌عنوان یک گزینهٔ انحرافی قانع‌کننده عمل کند.
توجه: فقط ضرب‌المثل‌های شناخته‌شده و معتبر انگلیسی را انتخاب کنید.
فرمت خروجی:

```
{
  "equivalent_proverb": "ضرب‌المثل انگلیسی معادل",
  "distractor_proverb": "ضرب‌المثل انگلیسی انحرافی"
}
```

<proverb>عاشقی کار سری نیست که بر بالین است.</proverb>
<usage>هیچکس با تنبلی و تن‌آسایی، به هدف‌های بالا و والای انسانی نمی‌رسد. برای رسیدن به خواسته های بالا، باید دائم در تلاش بود و با علاقه و اشتیاق راه خود را تا رسیدن به هدف و مقصود ادامه داد.</usage>

English Translation

You are a language model designed to suggest English proverbs that correspond to Persian proverbs.
You are given a Persian proverb along with its usage context.
Your task is to provide two real English proverbs:
1. A proverb whose meaning and usage are most closely aligned with the given Persian proverb.
2. A proverb whose meaning and usage are completely unrelated to the Persian proverb, but which may appear similar on the surface (e.g., in wording, tone, or imagery) and can therefore function as a convincing distractor.
Note: Only select well-known and established English proverbs.
Output format:

```
{
  "equivalent_proverb": "Equivalent English proverb",
  "distractor_proverb": "Distractor English proverb"
}
```

<proverb>Love is not a secret affair to be pursued on a sickbed.</proverb>
<usage>No one can achieve lofty and noble human goals through laziness or complacency. Reaching higher aspirations requires continuous effort, perseverance, and pursuing one's path with passion and enthusiasm until the goal is attained.</usage>

Figure 5: Example prompt for generating English equivalents of Persian proverbs for cross-cultural understanding, along with its English translation.

شما یک مدل زبانی هستید که برای تشخیص مفهوم ضرب‌المثل‌های فارسی در مکالمه طراحی شده‌اید.
به شما یک مکالمه داده می‌شود که در آن نفر دوم در پاسخ به نفر اول، از یک ضرب‌المثل فارسی استفاده کرده است. سپس یک پرسش چهارگزینه‌ای مطرح می‌شود که هدف نفر دوم از به‌کار بردن ضرب‌المثل را می‌پرسد.
وظیفه شما این است که تنها یک گزینه‌ی صحیح را انتخاب کنید. پاسخ شما باید فقط شامل حرف گزینه‌ی درست (الف، ب، ج یا د) باشد و هیچ متن یا توضیح اضافی ارائه نکنید.
مکالمه:
سارا: از بهرام چه خبر؟ قرار بود تو پیدا کردن خونه کمکمون کنه.
مریم: بی‌خیال بابا. اون صد تا چاقو بسازه، یکیش دسته نداره.
سؤال:
مریم با این ضرب‌المثل چه منظوری داشت؟
گزینه‌ها:
الف) بهرام آدم خطرناک و دعوایی و بهتره ازش دوری کنیم.
ب) بهرام چند تا کار رو با هم شروع می‌کنه ولی هیچکدوم رو به آخر نمی‌رسونه.
ج) بهرام فقط حرف می‌زنه و قول زیاد می‌ده، ولی به هیچکدوم از قول‌هاش عمل نمی‌کنه.
د) بهرام شغلش چاقوسازیه ولی اصلا حرفه‌ای نیست و چاقوهاش دسته ندارن.

English Translation

You are a language model designed to identify the intended meaning of Persian proverbs used in conversation.
You are given a dialogue in which the second speaker responds to the first speaker by using a Persian proverb. Then, a multiple-choice question is asked about the intention behind the second speaker's use of the proverb.
Your task is to select only one correct option. Your answer must consist only of the letter of the correct option (A, B, C, or D), and you must not provide any additional text or explanation.
Dialogue:
Sara: Any news about Bahram? He was supposed to help us find a place.
Maryam: Forget it. Even if he makes a hundred knives, none of them has a handle.
Question:
What did Maryam mean by using this proverb?
Options:
A) Bahram is a dangerous and aggressive person, and it is better to stay away from him.
B) Bahram starts many tasks at the same time but never finishes any of them.
C) Bahram only talks and makes many promises, but he does not keep any of them.
D) Bahram's job is knife-making, but he is not professional and his knives have no handles.

Figure 6: Example prompt for evaluating LLMs' contextual understanding of Persian proverbs, along with its English translation.

شما یک مدل زبانی هستید که برای تشخیص معادل معنایی مناسب ضرب‌المثل‌های فارسی در زبان انگلیسی طراحی شده‌اید.
به شما یک ضرب‌المثل فارسی داده می‌شود. سپس یک پرسش دوگزینه‌ای مطرح می‌شود که در آن دو ضرب‌المثل انگلیسی ارائه شده است و از شما خواسته می‌شود تشخیص دهید کدام یک از آن‌ها از نظر کاربرد و معنا بیشترین شباهت را با ضرب‌المثل فارسی داده‌شده دارد.
وظیفه‌ی شما این است که تنها یک گزینه‌ی صحیح را انتخاب کنید. پاسخ شما باید فقط شامل حرف گزینه‌ی درست (الف یا ب) باشد و هیچ متن یا توضیح اضافی ارائه نکنید.
ضرب‌المثل فارسی:
جایی که گوشت نیست، چغندر پهلوان است.
سؤال:
کدام یک از این دو ضرب‌المثل انگلیسی از نظر کاربرد و معنا بیشترین شباهت را با ضرب‌المثل فارسی داده‌شده دارد؟
گزینه‌ها:
الف) One man's meat is another man's poison.
ب) In the land of the blind, the one-eyed man is king.

English Translation

You are a language model designed to identify the most appropriate semantic equivalent of Persian proverbs in the English language.
You are given a Persian proverb. Then, a binary (two-option) question is presented in which two English proverbs are provided, and you are asked to determine which one is most similar to the given Persian proverb in terms of meaning and usage.
Your task is to select only one correct option. Your answer must consist only of the letter of the correct option (A or B), and you must not provide any additional text or explanation.
Persian proverb:
Where there is no meat, the beetroot is a champion.
Question:
Which of the following two English proverbs is most similar in meaning and usage to the given Persian proverb?
Options:
A) One man's meat is another man's poison.
B) In the land of the blind, the one-eyed man is king.

Figure 7: Example prompt for evaluating LLMs' cross-cultural understanding of Persian proverbs, along with its English translation.