TARAZ: Persian Short-Answer Question Benchmark for Cultural Evaluation of Language Models

Reihaneh Iranmanesh*, Saeedeh Davoudi*, Pasha Abrishamchian, Ophir Frieder, Nazli Goharian
IR Lab, Computer Science Department
Georgetown University, Washington D.C.
{rei, saeedeh, pasha, ophir, nazli}@ir.cs.georgetown.edu

Abstract

This paper presents a comprehensive evaluation framework for assessing the cultural competence of large language models (LLMs) in Persian. Existing Persian cultural benchmarks rely predominantly on multiple-choice formats and English-centric metrics that fail to capture Persian's morphological complexity and semantic nuance. Our framework introduces a Persian-specific short-answer evaluation that combines rule-based morphological normalization with a hybrid syntactic and semantic similarity module, enabling robust soft-match scoring beyond exact string overlap. Through systematic evaluation of 15 state-of-the-art open- and closed-source models across three culturally grounded Persian datasets, we demonstrate that our hybrid evaluation improves scoring consistency by +10% compared to exact-match baselines by capturing meaning that surface-level methods cannot detect. Our human evaluation further confirms that the proposed semantic similarity metric achieves higher agreement with human judgments than LLM-based judges. We publicly release our evaluation framework, providing the first standardized benchmark for measuring cultural understanding in Persian and establishing a reproducible foundation for cross-cultural LLM evaluation research.

Keywords: Persian NLP, cultural evaluation, large language models, short-answer questions

1. Introduction

Large Language Models (LLMs) excel at general linguistic tasks (Brown et al., 2020; OpenAI, 2025), yet struggle with cultural common sense understanding (Shen et al., 2024). Despite growing interest in cross-cultural evaluation (Arora et al., 2025; Li et al., 2024b,a), Persian remains notably absent from these major cultural benchmarks.

Few recent studies on the Persian language (Saffari et al., 2025; Moosavi Monazzah et al., 2025) have explored incorporating cultural understanding in Multiple Choice Questions (MCQs) but recent research (Chamieh et al., 2024) has shown that MCQ task assesses factual recall rather than conceptual understanding. Short-Answer Question (SAQ) tasks require free-form responses, providing a more direct probe of cultural and semantic understanding (Saffari et al., 2025). However, SAQ datasets and evaluation benchmarks that explicitly account for Persian cultural understanding remain limited.

BLEnD (Myung et al., 2024) is the first benchmark on SAQ task that is multilingual covering 13 languages including Persian (known as Farsi), providing opportunity for further research. Meanwhile, BLEnD suffers from culturally misaligned examples (Moosavi Monazzah et al., 2025). Their exact match evaluation metric cannot capture morphological Persian variations that express the same semantic content. For example, نان ("bread") versus نون ("bread," informal) are semantically identical but syntactically different. While some works employ LLM-based graders as an alternative, research demonstrates that these models exhibit instability, prompt sensitivity, and positional bias. (Luo et al., 2023; Lee et al., 2025; Farzi and Dietz, 2024).

To address these challenges, we present TARAZ, a comprehensive evaluation framework that assesses LLMs' cultural competence in Persian through short-answer questions task. Our contributions are:[1]

- ISN-SAQ and PerCul-SAQ: SAQ adaptation of the Iranian Social Norms (ISN) classification dataset (Saffari et al., 2025) and PerCul MCQ dataset (Moosavi Monazzah et al., 2025), transforming 1127 examples into culture-specific questions that test normative social behaviors across diverse environmental and demographic contexts.
- Persian-specific LLMs' responses postprocessing, handling morphological and numeric variation (textual, and symbolic forms across Persian/Arabic numeral systems) and conjunction-based segmentation.
- A hybrid syntactic and semantic-based evaluation to achieve robust semantic matching instead of exact matching. Our evaluation framework is extensible to new datasets and any language model's responses.

*Equal contribution.

[1] Code, data, and evaluation are available at TARAZ.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5256–5267
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

- Comprehensive evaluation of 15 state-of-the-art (multilingual and Persian) LLMs across six metrics on BLEnD, ISN-SAQ and PerCul-SAQ datasets.

In the rest of this paper, we first review the background and related works in Section 2. Section 3 presents the datasets used in our study, followed by Section 4, which describes the models employed. Section 5 details our evaluation framework, and Section 6 reports the experimental setup and results. Finally, Section 7 concludes the paper.

2. Related Work

In the following section, we provide a review of studies related to the evaluation of LLMs in Persian and cultural settings.

2.1. Persian NLP Benchmarks and Evaluation Challenges

Persian NLP has advanced through development of diverse benchmarks. ParsiNLU (Khashabi et al., 2021) provided early task coverage with 2.4K multiple-choice questions across literature, commonsense, and mathematics domains. ParsBench (Shariati Motlagh, 2025) established standardized evaluation protocols and developed a leaderboard ranking Persian LLMs on tasks including sentiment analysis, machine translation, and multiple-choice question answering. The Khayyam Challenge (Ghahroodi et al., 2024) adapted MMLU with 20,192 four-choice questions across 38 domains extracted from Persian examinations, introducing content spanning elementary to secondary education levels. These efforts employ MCQ formats, which measure factual recall and perform exact matching rather than comprehension (Luo et al., 2023).

Hosseinbeigi et al. (Hosseinbeigi et al., 2025a)[2] introduced PeKA (3,600 MCQs on Persian cultural knowledge) and PK-BETS (4,000 questions including open-ended generation tasks). However, their open-ended evaluation relies on LLM-as-a-Judge (GPT-4o), achieving only moderate agreement with human judgment (Cohen's κ=0.54), introducing evaluation bias, proprietary model dependency, and scalability limitations.

FarsEval-PKBETS (Hashemi et al., 2025)[2] introduced 3500 multiple-choice, and 500 short-answer questions covering medicine, law, religion, and Persian language tasks. However, their evaluation of short-answer questions relies on manual human annotation with subjective Correct/Wrong/Semi-correct labels, limiting scalability and potentially introducing inconsistency. They evaluated only three models and provided no automated evaluation framework for handling morphological variation in Persian SAQs. BLEnD (Myung et al., 2024) is introduced as the first open-source SAQ and MCQ benchmark that includes Farsi. It provides both datasets and evaluation metrics for the SAQ task. The BLEnD dataset is discussed in detail in Section 3.

2.2. Multilingual and Cultural-Linguistic Alignment

Ying et al.(Ying et al., 2025) showed that multilingual models perform better when the question's language aligns with its cultural context. CulFiT (Feng et al., 2025) introduces multilingual critique data synthesis with fine-grained rewards, achieving state-of-the-art open-source performance but still faces SAQ evaluation challenges. Evaluating models only in English uses translation-based methods that do not show real understanding of the target language. For example, cultural concepts like Taarof, a kind of polite social ritual in Farsi, have no clear English equivalent. TaarofBench (Sadr et al., 2025) confirms this pattern with dramatic improvements from English to Persian prompts: DeepSeek V3 improved by 32.0 points, GPT-4o by 33.1 points, and Claude 3.5 by 25.2 points. Matina (Hosseinbeigi et al., 2025b) introduces a multi-expert approach based on Low-Rank Adaptation (LoRA) to better align Persian text generation with cultural and linguistic nuances[3].

Our Persian-specific postprocessing (numeric normalization, conjunction segmentation, morphological normalization) addresses the evaluation gaps that prior work (Hashemi et al., 2025; Feng et al., 2025; Myung et al., 2024) could not resolve through manual effort alone, providing an automated and scalable solution for robust Persian SAQ evaluation.

2.3. LLM-Based Evaluation and Its Limitations

Chamieh et al.(Chamieh et al., 2024) demonstrate that zero-shot and few-shot LLM graders underperform supervised baselines. Luo et al.(Luo et al., 2023) show that even calibrated rubric prompting achieves only modest human agreement, typically plateauing at 70–80% regardless of prompt engineering effort. The EMBER benchmark (Lee et al., 2025) reveals that LLM judges systematically penalize epistemic uncertainty markers, re-

[2] Their dataset and evaluation code are not publicly available as of this submission.

[3] Their model is not publicly available as of this submission.

ducing scoring fairness for cautious but correct responses. The EXAM++ framework (Farzi and Dietz, 2024) documents technical pitfalls including positional biases (15–20% preference for first-shown answers) and brittle answer verification where minor formatting differences cause evaluation failure. Traditional SAQ approaches (Myung et al., 2024) rely on string-matching metrics (exact match, soft match) which fail to recognize semantically equivalent but syntactically distinct answers–a problem exacerbated in morphologically rich languages like Persian.

3. Datasets

We evaluate LLMs on three Persian cultural datasets, each targeting distinct aspects of cultural understanding. All datasets use native Persian prompts and references to ensure authentic evaluation without translation artifacts.

BLEnD The BLEnD (Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages) (Myung et al., 2024) dataset provides everyday cultural questions across 13 languages and 16 regions. We utilize the Persian data containing 500 questions spanning topics like food, sports, family, education, holidays/celebrations/leisure, and work-life. Each question has multiple human-annotated acceptable answers with associated frequency counts, enabling weighted evaluation. Questions are designed to probe implicit cultural knowledge that locals take for granted but may be absent from multilingual training corpora. We exclude questions where annotators predominantly responded with "I don't know" (idks $\geq$ 5) or "not applicable" responses ($\geq$ 3), following the original BLEnD paper. See Figure 4 for a culturally-specific example.

PerCul-SAQ The PerCul (Persian Cultural Understanding) (Moosavi Monazzah et al., 2025) dataset evaluates story-driven cultural reasoning through 592 multiple-choice questions derived from Persian cultural scenarios. Each entry consists of a Persian narrative describing a cultural situation, followed by a question with four answer choices. We introduce PerCul-SAQ, converting PerCul to short-answer format by changing the prompt, creating a more challenging generation task than the original multiple-choice setting. Stories contextualize cultural elements within realistic scenarios, requiring models to infer appropriate behavior or identify culturally-significant items rather than merely recalling facts.

The dataset spans 11 cultural categories: foods (خوراکی 191 questions), visible behavior (رفتار 56), iconic figures (چهره نمادین 55), institutions (نهاد 43), architecture (معماری 43), objects (شی 42), appropriacy (هنجار 36), dress (پوشاک 33), music (موسیقی/ساز 32), art (هنر 32), and rituals (آیین 29).

See Figure 5 for an example where a Nowruz celebration story requires identifying the Tombak (تنبک), a single-headed goblet drum central to Persian musical tradition.

ISN-SAQ We derive our ISN-SAQ dataset from the Persian Social Norms (PSN) dataset (Saffari et al., 2025), which contains 1,699 Iranian social norms with environment, context, and cultural appropriateness labels (Expected/Normal/Taboo). We convert norms labeled as Expected or Normal into short-answer questions, filtering out general norms and duplicates, yielding 535 culture-specific questions.

The questions cover diverse social situations: airport etiquette (e.g., در فرودگاه‌های ایران قبل از سفر چه رفتاری رایج است؟ / "What behavior is common at Iranian airports before travel?" → طلب حلالیت / requesting forgiveness), marketplace practices (در بازار ایران هنگام خرید با فروشندگان چه رفتاری رایج است؟ → چانه‌زنی / bargaining), family traditions (بعد از تولد پسران در ایران چه مراسمی برگزار می‌شود؟ → ختنه سوران / circumcision ceremony), and gender-specific behaviors. ISN-SAQ emphasizes situated social norms tied to specific environments (مسجد/mosque, بازار/marketplace, فرودگاه/airport, خانه/home) and demographic contexts, reflecting how appropriate behavior varies across settings within Iranian culture.

Figure 6 illustrates how a Bazaar norm (چانه‌زنی, bargaining) is reformulated as an SAQ, capturing context-dependent social behavior specific to Iranian marketplace culture.

Table 1 summarizes dataset characteristics. BLEnD emphasizes everyday factual knowledge with multiple valid answers per question, PerCul-SAQ tests cultural narrative reasoning with single correct answers, and ISN-SAQ focuses on normative social behaviors with highly specific contextual requirements. BLEnD, PerCul-SAQ, and ISN-SAQ each target a single facet of Persian cultural competence – everyday facts, story-based reasoning, and social norms respectively. No prior work evaluates all three jointly or provides a unified SAQ scoring framework applicable across them. Our work, TARAZ, fills this gap.

4. Models

We evaluate 15 generative models spanning three categories: closed-source proprietary models, open-weight foundation models, and Persian fine-tuned models. Since SAQ evaluation requires text generation, we focus on Generative Language Models and exclude encoder-based models (e.g.,

	BLEnD	PerCul-SAQ	ISN-SAQ
Questions	500	592	535
Answers/Q	Multi (3.7)	Single	Single
Avg. Q len.	11.0	99.5	9.1
Avg. A len.	1.6	3.1	2.7
Focus	Knowledge	Stories	Norms

Table 1: Comparison of short-answer question (SAQ) datasets. *Questions* indicates the total number of questions; *Answers/Q* shows the number of valid answers per question; *Avg. Q len.* and *Avg. A len.* denote average question and answer lengths in words (whitespace-tokenized); and *Focus* describes each dataset's main content type. Note that PerCul-SAQ questions are narrative scenarios rather than direct questions, explaining the longer average length.

ParsBERT (Farahani et al., 2021), which are designed for classification rather than open-ended response generation). Persian models were selected based on their ranking across three major Persian LLM leaderboards: MIZAN (MCINext Team, 2025), ParsBench (Shariati Motlagh, 2025), and Open Persian LLM Leaderboard (PartAI and AUT NLP Lab, 2024).

4.1. Closed Source Models

We evaluate four frontier proprietary models with API-based access: *GPT-5* (OpenAI, 2025), *Claude Opus 4.1* (Anthropic, 2025a), *Claude Sonnet 4.5* (Anthropic, 2025b), and *Gemini 2.5 Flash-Lite* (Comanici et al., 2025).

4.2. Open Weight Models

We evaluate five open-weight foundation models: *Qwen2.5-72B* (Qwen et al., 2025) (multilingual with explicit Persian support), *LLaMA 3.1-70B* (Touvron et al., 2023), *DeepSeek-V3* (DeepSeek-AI et al., 2025) (mixture-of-experts architecture), *Gemma-2-27B* (Team et al., 2024), and *Phi-4-14B* (Abdin et al., 2025).

4.3. Persian Fine-Tuned Models

We evaluate six Persian-adapted models representing diverse adaptation strategies: *PersianMind* (Rostami et al., 2024) extends LLaMA2-7B with 10,000 Persian tokens and trains on 2B Persian tokens; *ParsT5* (Puraminy, 2021) adapts T5 encoder-decoder through masked language modeling on 35GB Persian text; *Dorna* (PartAI, 2024) fine-tunes LLaMA 3-8B-Instruct on Persian instructions; *PersianLLaMA* (Abbasi et al., 2023) (13B) trains adapters on pretrained LLaMA2; *Maral* (Haghiri and Mohrechi, 2023) fine-tunes Mistral-7B on machine-translated Alpaca Persian; and *AVA-LLAMA* (Moghadam, 2024) applies parameter-efficient fine-tuning to LLaMA 3-8B.

5. Evaluation

5.1. Evaluation Metrics

To evaluate the capabilities of different language models on Persian cultural knowledge, we use syntactic metrics (Exact Match and ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004)), semantic-based approaches including GPT-5 as a judge, and cosine similarities using Transformer-based Persian embeddings. All metrics assess responses of LLMs against ground truth data. At the end, all techniques assign a binary score to each question–answer pair: “1” if the answer is correct and “0” otherwise. The definitions of each metric are provided below:

1. **Exact Match (EM)**: Normalized Responses are lemmatized using the BLEnD benchmark. A response is labeled as ”1” as long as it exactly matches one of the ground-truth data.

2. **ROUGE Metrics:** We employ three ROUGE metrics (Lin, 2004): (1) *ROUGE-1* measures unigram overlap between the model response and ground truth; (2) *ROUGE-2* measures bigram overlap (consecutive word pairs); and (3) *ROUGE-L* calculates the longest common subsequence. All scores are computed as F1 measures. If the F1 score is 1.0, we label the response as ”1”; otherwise, it will be labeled as ”0”.

3. **LLM-judge:** GPT-5 evaluates semantic equivalence of model responses against ground truth answers using the following prompt:

```
text = f"""
You are evaluating a model's answer.

Question ID: {qid}
Response: "{response}"

Valid answers: {answers}

Does the response semantically match
ANY of the valid answers?
Answer ONLY with 1 if it matches,
0 if it does not.
"""
```

   If GPT-5 returns 1, the answer is labeled as “1”; otherwise, it is labeled as “0.”

4. **Cosine similarity using a Transformer-based model.** *maux-gte-persian-v3* (xmanii,

2024) represents a high-performance Persian sentence embedding solution fine-tuned from Alibaba-NLP/gte-multilingual-base (Zhang et al., 2024). This model produces 768-dimensional sentence embeddings with support for sequences up to 8192 tokens, making it suitable for long-form Persian text processing. Responses with cosine similarity above an empirically determined threshold 0.85 are classified as "1", and "0" otherwise. Using this model, we are able to generate direct sentence-level embeddings, allowing for more sophisticated semantic understanding. The system calculates the maximum sentence similarity between any sentence in the model response and any ground truth annotations.

5. **Hybrid approach:** We applied Persian-specific postprocessing (Section 5.2) on each model's responses before computing cosine similarity using *maux-gte-persian-v3* embeddings. In that way, we could handle numeric normalization (e.g., ۱۲۳ vs صد و بیست و سه), conjunction-based segmentation, morphological normalization, and diacritic removal that were not handled by previous benchmarks. The hybrid method is the most robust metric for Persian evaluation (Section 5.2).

5.2. Postprocessing

Persian presents persistent challenges for natural language processing due to its morphological complexity, orthographic ambiguity, and inconsistent writing conventions. The absence of clear word boundaries, where clitics and compound words are often written without spaces, complicates tokenization and lemmatization. Orthographic inconsistencies, such as optional half-spaces and variable diacritic use, result in multiple valid spellings of the same word, for example, می رود ("goes") versus میرود ("goes"). Similarly, lexical variation between formal and colloquial registers creates normalization challenges, as in نان ("bread") versus نون ("bread," informal). Moreover, Persian text often mixes symbols, numbers, and characters from different writing systems including multiple numeral encodings (e.g., ۶ vs. 6, ششم vs شش vs شیش), inconsistent use of Arabic and Persian letters (e.g., یک vs يك), and diverse representations of dates (e.g., ۱۴۰۰ / ۰۷ / ۲۵ → "the twenty-fifth of Mehr, fourteen hundred") and times (e.g., ۱۱:۳۵ → "eleven thirty-five"). These examples illustrate how surface variability and non-standardized writing conventions introduce substantial ambiguity, making Persian an especially demanding language for robust text normalization and evaluation. During postprocessing, we address these issues to ensure consistent and comparable evaluation of Persian texts:

- *Whitespace and punctuation normalization:* Removes extra spaces and trailing punctuation marks such as ".", "،", "؛", "!", and "؟".
- *Digit normalization:* Converts both English (0-9) and Arabic (٠ - ٩) numerals into their Persian equivalents (۰ - ۹) (e.g., 56 vs ۵۶ vs ٥٦).
- *Character normalization:* Replaces Arabic variants of certain letters with their Persian counterparts, for example:
  - ي → ی
  - ك → ک
  - ة → ه
  - Removes diacritics and redundant ء.
- *Diacritic removal:* Eliminates Persian vowel marks to standardize text.
- *Stop word removal:* Removes common Persian stop words using the `hazm` library, keeping only meaningful tokens.
- *Suffix removal (stemming):* Strips frequent Persian suffixes such as ات, یات, ها, ان, ین, ون, and گان to reduce words to their base form.
- Splits text using Persian and English punctuation, including semicolons (؛, ;) and commas (،, ,).
- Further splitting items by conjunctions when they occur *between two substantial parts*:
  - و (and)
  - یا (or)
  - هم (also)
  - همینطور (likewise)
  - نیز (also)
  - همچنین (in addition)
- Removes any empty or redundant entries, returning a list of clean, distinct response items.

6. Experiments and Results

Our experimental plan evaluates LLMs on their ability to understand and generate culturally grounded responses in Farsi using short-answer questions (SAQs). Evaluation is conducted on three datasets: BLEnD, PerCul-SAQ, ISN-SAQ, which together cover everyday culture, narratives, and normative social behaviors. Each model receives only Farsi prompts to ensure evaluation within native linguistic context. We report Exact Match, ROUGE-1, ROUGE-2, ROUGE-L, GPT-5

	Model	EM	ROUGE-1	ROUGE-2	ROUGE-L	LLM-judge	Maux	Maux+Post
Closed Source	GPT-5	73.246	42.105	12.939	42.105	71.930	84.652	85.837
	Claude-Opus-4.1	79.386	44.737	11.184	44.739	81.579	84.869	85.781
	Claude-Sonnet-4.5	74.561	42.982	9.868	42.763	78.509	85.307	86.764
	Gemini-2.5-Flash-Lite	54.605	34.649	11.404	21.711	58.772	62.061	62.854
Open Weight	LLaMA-3.1-70B-Inst	47.368	21.491	6.140	21.491	57.018	63.158	64.493
	DeepSeek-V3	43.421	21.711	5.921	42.213	45.833	48.247	50.012
	Gemma-2-27B-IT	70.614	37.500	8.991	37.500	74.123	78.947	79.391
	Phi-4-14B	25.000	5.263	0.439	5.263	23.465	72.012	73.204
	Qwen-2.5-72B-Instruct	53.289	23.465	4.386	23.246	31.140	73.521	74.812
Persian Fine-Tuned	PersianMind-v1.0-7B	32.018	4.605	1.316	4.600	25.000	27.227	28.018
	ParsT5	1.974	0.219	0.219	4.005	0.877	10.151	10.174
	Dorna-LLaMA3-8B-Instruct	49.123	20.395	3.070	20.395	47.807	54.824	55.064
	PersianLLaMA-13B	21.930	15.965	1.681	15.967	0.000	5.204	6.214
	Maral-7B	18.860	9.211	1.535	8.991	20.175	21.181	22.229
	AVA-LLaMA-3-8B	49.781	30.711	3.006	34.429	39.333	42.359	58.647

Table 2: Performance comparison of closed-source, open-weight, and Persian fine-tuned language models on the BLEnD dataset. Reported metrics include Exact Match (EM), ROUGE-1, ROUGE-2, and ROUGE-L for lexical similarity; LLM-judge, Maux and Maux+Post for semantic evaluation. Closed-source models (e.g., GPT-5, Claude) outperform others across most metrics, while open-weight models like Gemma-2 and fine-tuned Persian models (e.g., AVA-LLAMA3-8B) demonstrate competitive performance within their respective categories.

	Model	EM	ROUGE-1	ROUGE-2	ROUGE-L	LLM-judge	Maux	Maux+Post
Closed Source	GPT-5	45.946	29.899	18.750	29.899	74.493	79.493	80.493
	Claude-Opus-4.1	39.358	29.899	16.047	29.899	60.304	65.304	66.304
	Claude-Sonnet-4.5	38.345	25.338	13.176	25.338	57.770	62.770	63.770
	Gemini-2.5-Flash-Lite	19.426	15.372	7.264	15.203	35.473	40.473	41.473
Open Weight	LLaMA-3.1-70B-Inst	12.162	7.264	2.703	7.095	21.622	26.622	27.622
	DeepSeek-V3	30.912	23.986	11.824	23.818	48.649	53.649	54.649
	Gemma-2-27B-IT	17.905	14.696	8.277	14.527	30.236	35.236	36.236
	Phi-4-14B	1.182	0.000	0.000	0.000	4.223	9.223	10.223
	Qwen-2.5-72B-Instruct	3.716	2.365	1.182	2.365	17.230	22.230	23.230
Persian Fine-Tuned	PersianMind-v1.0-7B	–	0.000	0.000	0.000	23.142	28.142	29.142
	ParsT5	0.000	0.000	0.000	0.000	1.689	6.689	7.689
	Dorna-LLaMA3-8B-Instruct	7.284	0.845	0.189	0.845	9.459	14.459	15.459
	PersianLLaMA-13B	0.845	0.000	0.000	0.000	20.101	25.101	26.101
	Maral-7B	–	0.000	0.000	0.000	33.446	38.446	39.446
	AVA-LLaMA-3-8B	5.068	0.000	0.000	0.000	6.588	11.588	12.588

Table 3: Performance of closed-source, open-weight, and Persian fine-tuned models on the PerCul-SAQ dataset, which evaluates story-based cultural reasoning across 11 cultural categories. Metrics include Exact Match (EM), ROUGE-1, ROUGE-2, and ROUGE-L for lexical overlap; LLM-judge for semantic equivalence; and Maux and Maux+Post for human-aligned cultural judgment. Closed-source models achieve higher overall scores, while fine-tuned Persian models show low capability in culturally grounded comprehension tasks.

as LLM-judge, Maux, and Maux+Post metrics. All evaluations are performed in a zero-shot setting with fixed temperature (T = 0.2) if applicable, and outputs are normalized before scoring. Finally, we analyze results across cultural domains and model types to further understand the strengths and weaknesses of each model.

Category	Model	EM	ROUGE-1	ROUGE-2	ROUGE-L	LLM-judge	Maux	Maux+Post
Closed Source	GPT-5	20.187	8.224	4.673	6.224	41.682	46.682	46.682
	Claude-Opus-4.1	20.374	8.037	3.925	7.664	42.056	47.056	47.056
	Claude-Sonnet-4.5	18.131	7.684	3.925	7.477	37.757	42.757	42.757
	Gemini-2.5-Flash-Lite	16.636	9.346	4.299	3.346	33.084	38.084	38.084
Open Weight	LLaMA-3.1-70B-Inst	11.589	6.729	3.364	6.729	27.290	32.290	32.209
	DeepSeek-V3	17.944	4.120	1.301	4.120	44.299	49.299	49.229
	Gemma-2-27B-IT	13.271	6.729	2.991	6.729	32.150	37.150	37.150
	Phi-4-14B	3.364	0.374	0.000	0.374	9.346	14.346	14.346
	Qwen-2.5-72B-Instruct	9.346	1.495	0.935	1.495	31.776	36.776	36.764
Persian Fine-Tuned	PersianMind-v1.0-7B	3.178	2.056	1.308	2.056	8.224	14.224	14.224
	ParsT5	0.187	0.000	0.000	0.000	0.561	5.561	6.561
	Dorna-LLaMA3-8B-Instruct	4.299	1.121	0.748	1.121	15.327	20.327	20.327
	PersianLLaMA-13B	2.617	0.000	0.000	0.000	5.981	10.981	10.981
	Maral-7B	0.374	0.000	0.000	0.000	1.869	6.869	6.869
	AVA-LLaMA-3-8B	4.486	1.121	0.748	1.121	15.701	20.701	20.701

Table 4: Model performance on the ISN-SAQ dataset across key metrics (EM, ROUGE, LLM-judge, and Maux scores). Maux+Post metric achieves the strongest results overall, with Claude Opus and DeepSeek-V3 leading. Persian fine-tuned models perform notably lower, the same pattern we observed in previous datasets.

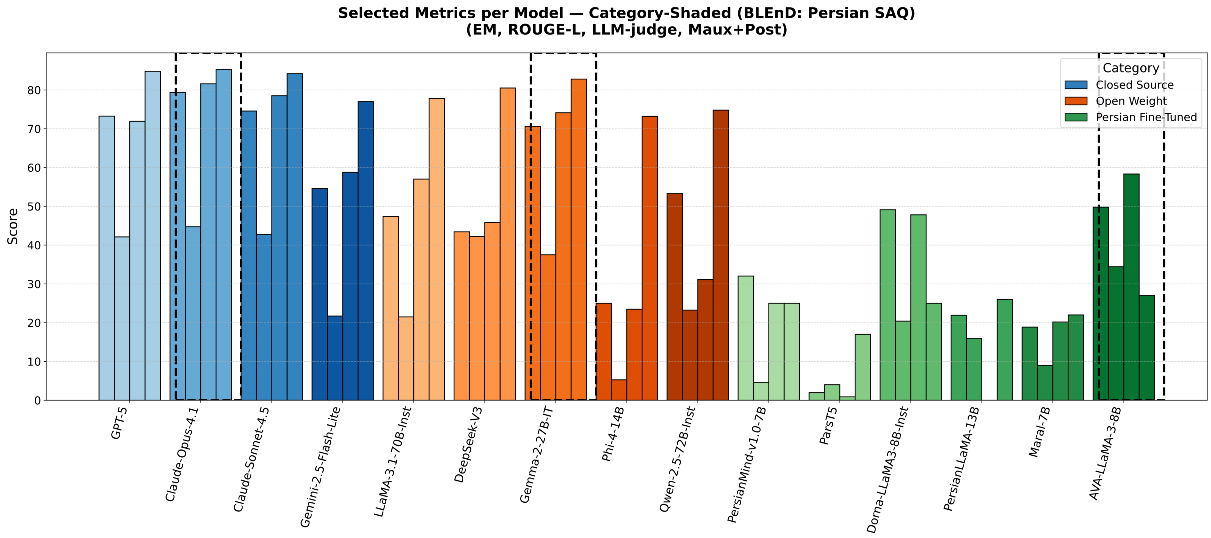

Figure 1: Comparison of model performance on the BLEnD dataset using four key metrics (left to right) – Exact Match (EM), ROUGE-L, LLM-judge, and Maux+Post. Bars are grouped per model and color-coded by category (Closed Source, Open Weight, Persian Fine-Tuned). Each dashed box highlights the best-performing model within its category, based on normalized average performance across all selected metrics. Claude-Opus-4.1 is the best closed-source model, Gemma-2-27B-IT is the best open-weight model and AVA-LLaMA-3-8B is the best Persian fine-tuned model. Closed-source models achieve the highest scores overall.

Overall Performance Closed-source models perform best across all datasets (Tables 2, 3, 4). Claude-Opus-4.1 and GPT-5 reach the highest Maux+Post scores, while Gemma-2-27B and DeepSeek-V3 are the strongest open-weight models. PerCul-SAQ and ISN-SAQ are harder than BLEnD, with large drops in lexical scores and semantic metrics. Persian fine-tuned models fall behind: Maral and PersianMind often produce repetitive or meaningless tokens, making their outputs effectively unusable as reflected in zero Rouge scores (See Tables 3, 4). Meanwhile, Gemma-2-27B outperforms all Persian-specific models despite no Persian-targeted fine-tuning.

Subtopic Analysis Based on Figure 1, we selected the top model across each model family

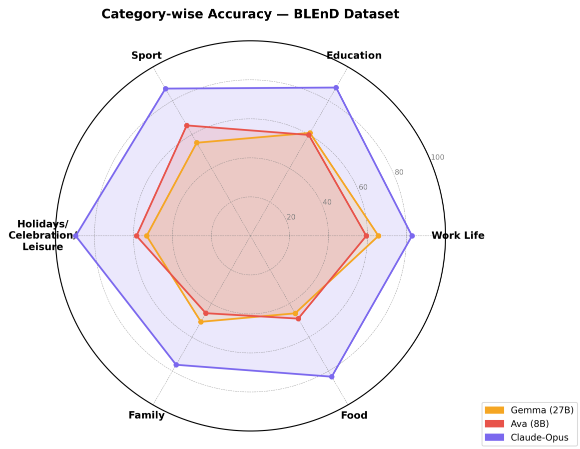

Figure 2: Category-wise accuracy on BLEnD. The plot shows accuracy for the top three models on BLEnD dataset. Claude Opus shows better accuracy among different categories compared to Gemma-2-27-IT and Ava-LLaMA-3-8B Persian model.

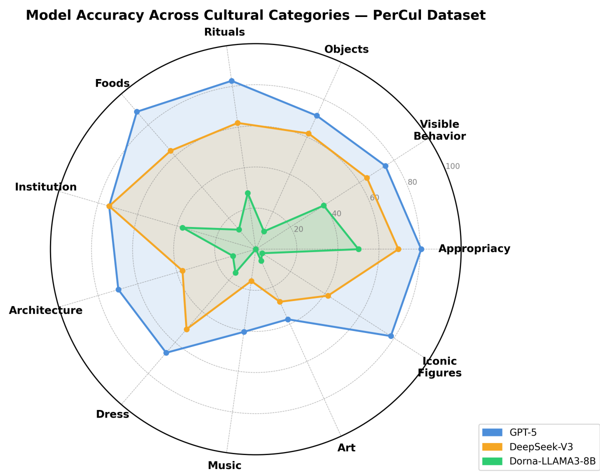

Figure 3: Category-wise accuracy on PerCul-SAQ dataset. The plot shows accuracy for the top three models among each closed-, open, and Persian models.

for subtopic analysis of BLEnD. As seen in Figure 2, on BLEnD, Claude Opus performs consistently above 76% across all subtopics, peaking at Holidays/Celebration/Leisure (89.6%) and Education (87.8%), while its weakest category, Family (76.3%), still surpasses both smaller models by a wide margin. Notably, Ava (8B) and Gemma (27B) exhibit similar performance profiles despite their substantial difference in parameter count, with overlapping scores across nearly every category: Gemma's best is Work Life (65.6%) and Ava's is Sport (65.4%). Both models struggle most with Food (Gemma: 45.8%, Ava: 49.0%) and Family (Gemma: 50.9%, Ava: 45.8%), pointing to persistent gaps in everyday cultural knowledge.

As seen in Figure 3, performance varies sharply across cultural subtopics on PerCul-SAQ. GPT-5 leads across all categories, with highest accuracy in Foods (88.5%) and Rituals (82.8%), and even its weakest categories – Art (37.5%) and Music (40.6%) – remain competitive. DeepSeek performs best on Institution (74.4%) and Appropriacy (69.4%), but drops sharply in Music (15.6%) and Art (28.1%). Dorna (8B) struggles severely across the board, with near-zero scores in Music (0%), Iconic Figures (3.6%), and Art (6.3%), and only marginally better performance in Appropriacy (50%) and Visible Behavior (39.3%), revealing that Persian cultural generalization remains far out of reach for smaller specialized models.

Rater Pair	Agreement (%)	Cohen's κ
LLM-judge vs. Annotator 1	77.9	0.015
LLM-judge vs. Annotator 2	83.8	0.461
LLM-judge vs. Annotator 3	80.9	0.271
Annotator 1 vs. Annotator 2	80.0	0.205
Annotator 1 vs. Annotator 3	85.7	0.300
Annotator 2 vs. Annotator 3	80.0	0.300
Maux vs. Annotator 1	93.4	-0.025
Maux vs. Annotator 2	85.2	0.157
Maux vs. Annotator 3	90.2	0.228
LLM-judge vs. human majority	82.4	0.244
Maux vs. human majority	95.1	0.384

Table 5: Inter-rater agreement between the LLM-judge, Maux, and three human annotators on 70 BLEnD samples. Percent agreement and Cohen's κ are reported for all pairwise combinations. Maux achieves higher agreement with human majority than LLM-judge across both metrics.

Human Evaluation Three annotators independently labeled 70 GPT-5 BLEnD responses as correct (1) or incorrect (0) to validate our automatic metrics. Table 5 reports percent agreement and Cohen's κ across all rater pairs. Human annotators agree 80.0–85.7% of the time with κ ranging 0.205–0.300, reflecting moderate agreement given the open-ended nature of the task. Low κ values across all pairs – including among humans – reflect a known prevalence artifact (Feinstein and Cicchetti, 1990): with over 85% of labels positive, κ is deflated by design and raw agreement is the more informative signal.

The LLM-judge achieves 77.9–83.8% agreement against individual annotators and 82.4% against the human majority ($\kappa = 0.244$), but is systematically stricter, producing 9 false negatives versus only 3 false positives – suggesting it penalizes surface-divergent but semantically correct responses. Maux outperforms the LLM-judge on both metrics, achieving 85.2–93.4% per-annotator agreement and 95.1% against the human majority ($\kappa = 0.384$). The negative κ for Maux vs. Annotator 1 ($\kappa =$-0.025) is an artifact of near-zero variance in that annotator's labels rather than genuine disagreement. Taken together, these results suggest that Maux better captures human judgment of semantic correctness for Persian cultural responses than the LLM-judge.

7. Conclusion

We present TARAZ, a comprehensive evaluation framework for assessing Persian cultural competence in large language models through short-answer questions. Our contribution addresses critical gaps in existing Persian NLP evaluation: while prior benchmarks rely on multiple-choice formats that permit pattern matching without genuine understanding, our framework requires open-ended generation and employs semantic similarity metrics robust to Persian's morphological complexity. Through systematic evaluation of 15 state-of-the-art models, we demonstrate that semantic similarity with Persian-specific postprocessing outperforms traditional lexical metrics. Our Persian-specific postprocessing pipeline handles numeric normalization, conjunction segmentation, and morphological variation which proves essential for robust evaluation, with consistent improvements across all model categories.

Our findings reveal that closed-source models consistently lead across all three datasets, with GPT-5 and Claude Opus achieving the strongest results on BLEnD and PerCul-SAQ, though all models – including frontier ones – struggle considerably on ISN-SAQ, revealing that Iranian social norm reasoning remains an open challenge. Among open-weight models, Gemma-2-27B performs remarkably well on BLEnD (EM: 70.6%), approaching closed-source levels, while DeepSeek-V3 shows the strongest open-weight performance on PerCul-SAQ and ISN-SAQ. Persian fine-tuned models lag far behind across all three datasets and all metrics, underscoring that language-specific fine-tuning alone is insufficient and that Persian cultural and social reasoning requires substantially more targeted training investment.

We publicly release all datasets, evaluation code, and postprocessing modules to establish a standardized benchmark for Persian cultural understanding and provide a reproducible foundation for cross-cultural LLM evaluation research.

Limitations

Our work has several limitations. We focus only on text-based evaluation and do not cover multimodal or vision-language understanding, which is an important direction for future research. Like all human-annotated datasets, ours may reflect annotator bias and fail to capture the full range of cultural interpretations across Persian-speaking regions and demographics.

Additionally, our framework is language-specific and tailored to Persian's morphology and cultural context. While the general method can be applied to other morphologically rich languages, the normalization and postprocessing steps would need to be redesigned.

Moreover, both Maux metrics depend on *maux-gte-persian-v3*, which is fine-tuned on formal Persian text and may not favor colloquial text. Bias analysis of this embedding model remains future work.

Finally, the poor performance of Persian fine-tuned models likely stems from three factors: training on machine-translated data introduces translationese artifacts; fine-tuning is predominantly applied to 7–8B parameter models, far below the 27–72B scale of stronger multilingual models; and instruction tuning optimizes format rather than injecting cultural knowledge. Future Persian LLMs should prioritize native instruction data and culturally grounded pretraining over language-specific adaptation alone.

Ethics Statement

This work involves human annotation of model outputs for evaluation purposes. Annotators were recruited from the research team and provided informed consent. The datasets used (BLEnD, PerCul, ISN) are publicly available and were collected by their respective authors under appropriate ethical guidelines. Our ISN-SAQ and PerCul-SAQ adaptations involve no new human subjects data collection. The cultural norms described in ISN-SAQ reflect documented Iranian social practices and do not represent the views of the authors. All models were accessed via official APIs or public repositories.

Bibliographical References

Mohammad Amin Abbasi, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, and Behrouz Minaei Bidgoli. 2023. Persianllama: Towards building first persian large language model.

Marah Abdin et al. 2025. Phi-4-reasoning technical report.

Anthropic. 2025a. System card addendum: Claude Opus 4.1. Technical report.

Anthropic. 2025b. System card: Claude Sonnet 4.5. Technical report.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2025. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11772–11817, Vienna, Austria. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Imran Chamieh, Torsten Zesch, and Klaus Giebermann. 2024. LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 309–315, Mexico City, Mexico. Association for Computational Linguistics.

Gheorghe Comanici et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

DeepSeek-AI, Aixin Liu, et al. 2025. Deepseek-v3 technical report.

Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. Parsbert: Transformer-based model for persian language understanding. *Neural Processing Letters*, 53(6):3831–3847.

Naghmeh Farzi and Laura Dietz. 2024. Exam++: Llm-based answerability metrics for ir evaluation. In *LLM4Eval: The First Workshop on Large Language Models for Evaluation in Information Retrieval*, pages 1–20, Washington DC, United States. CEUR Workshop Proceedings.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. 2025. Culfit: A fine-grained cultural-aware llm training paradigm via multilingual critique data synthesis.

Omid Ghahroodi, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. 2024. Khayyam challenge (persianmmlu): Is your llm truly wise to the persian language?

Muhammadreza Haghiri and Mahi Mohrechi. 2023. Maral-7B-alpha-1: A Persian large language model. `https://huggingface.co/MaralGPT/Maral-7B-alpha-1`.

Seyed Mohammad Hossein Hashemi, Zahra Vatankhah, Motahareh Ramezani, Niki Pourazin, Tara Zare, et al. 2025. Farsevalpkbets: A new diverse benchmark for evaluating persian large language models.

Sara Bourbour Hosseinbeigi, Behnam Rohani, Mostafa Masoudi, Mehrnoush Shamsfard, Zahra Saaberi, Mostafa Karimi Manesh, and Mohammad Amin Abbasi. 2025a. Advancing Persian LLM evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2711–2727, Albuquerque, New Mexico. Association for Computational Linguistics.

Sara Bourbour Hosseinbeigi, Fatemeh Taherinezhad, Heshaam Faili, Hamed Baghbani, Fatemeh Nadi, and Mostafa Amiri. 2025b. Matina: A large-scale 73B token Persian text corpus. 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 9143–9157, Albuquerque, New Mexico. Association for Computational Linguistics.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, et al. 2021. ParsiNLU: A suite of language understanding challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. 2025. Are LLM-judges robust to expressions of uncertainty? investigating the effect of epistemic markers on LLM-based evaluation. 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 8962–8984, Albuquerque, New Mexico. Association for Computational Linguistics.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 84799–84838. Curran Associates, Inc.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

MCINext Team. 2025. Mizan: A persian llm leaderboard. Hamrah Aval (MCI Next), hosted on Hugging Face.

Mehdi Hosseini Moghadam. 2024. AVA-LLama-3: Fine-tuned LLama 3 Persian large language model. https://github.com/mehdihosseinimoghadam/AVA-Llama-3.

Erfan Moosavi Monazzah, Vahid Rahimzadeh, Yadollah Yaghoobzadeh, Azadeh Shakery, and Mohammad Taher Pilehvar. 2025. PerCul: A story-driven cultural evaluation of LLMs in Persian. 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 12670–12687, Albuquerque, New Mexico. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages.

OpenAI. 2025. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/.

PartAI. 2024. Dorna-LLama3-8B-instruct. https://huggingface.co/PartAI/Dorna-Llama3-8B-Instruct.

PartAI and AUT NLP Lab. 2024. Open persian llm leaderboard. Developed by PartAI in collaboration with Amirkabir University of Technology (AUT) NLP Lab.

Ahmad Puraminy. 2021. ParsT5: A T5 model for Persian. https://huggingface.co/Ahmad/parsT5-base.

Qwen, An Yang, et al. 2025. Qwen2.5 technical report.

Pedram Rostami, Ali Salemi, and Mohammad Javad Dousti. 2024. Persianmind: A cross-lingual persian-english large language model.

Nikta Gohari Sadr, Sahar Heidariasl, Karine Megerdoomian, Laleh Seyyed-Kalantari, and Ali Emami. 2025. We politely insist: Your LLM must learn the Persian art of taarof. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1819–1838, Suzhou, China. Association for Computational Linguistics.

Hamidreza Saffari, Mohammadamin Shafiei, Donya Rooein, Francesco Pierri, and Debora Nozza. 2025. Can I introduce my boyfriend to my grandmother? evaluating large language models capabilities on Iranian social norm classification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6060–6074, Albuquerque, New Mexico. Association for Computational Linguistics.

Shahriar Shariati Motlagh. 2025. ParsBench: A toolkit for benchmarking Persian language models. https://github.com/ParsBench/ParsBench. GitHub repository.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. 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 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, et al. 2024. Gemma 2: Improving open language models at a practical size.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

xmanii. 2024. maux-gte-persian-v3: High-performance Persian sentence embedding model. https://huggingface.co/xmanii/maux-gte-persian-v3.

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling language and culture for evaluating multilingual large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22230–22251, Vienna, Austria. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, et al. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

Appendix A. Dataset Examples

ID: **A1-en-01** **Topic:** Food

QUESTION

یک میان وعده معمول برای بچه‌های پیش دبستانی در ایران چیست؟

(What is a common snack for preschool kids in Iran?)

ACCEPTED ANSWERS (WITH ANNOTATOR FREQUENCY)

Persian	English	Count
میوه / میوه	fruit	3
لقمه	sandwich	1
کیک و شیر	cake and milk	1
نون و پنیر	bread and cheese	1
تخم‌مرغ	eggs	1
کورنفلکس	cereal	1

Figure 4: An example from the BLEnD Persian subset. Multiple valid answers with annotator frequency counts reflect genuine variation in Iranian households – fruit (میوه) is the most common response, while bread and cheese (نون و پنیر), a staple of Iranian home meals, also appears. This diversity motivates weighted evaluation beyond exact match.

ID: **1_O_MUS** **Category:** ساز یا موسیقی

STORY

در یک شب بهاری، در یکی از روستاهای قدیمی ایران، مردم برای جشن نوروز گرد هم آمده بودند. صدای موسیقی از دور به گوش می‌رسید و همه با شور و شوق به سمت میدان اصلی روستا حرکت می کردند. در مرکز میدان، گروهی از نوازندگان با لباس‌های رنگارنگ و سنتی در حال نواختن بودند. یکی از نوازندگان با مهارت خاصی بر روی یک ساز کوبه‌ای می‌نواخت که صدای آن به خوبی با دیگر سازها هماهنگ بود و ریتمی شاد و پرانرژی به موسیقی می‌بخشید. مردم با شنیدن این صدا به وجد آمده و شروع به رقص و پایکوبی کردند. این ساز کوبه‌ای که در دست نوازنده بود، به نوعی روح جشن را به نمایش می گذاشت و همه را به حرکت وادار می کرد.

PROMPT (NEW)

با توجه به متن بالا چه ساز یا موسیقی در فرهنگ یا زندگی روزمره ایرانی قابل برداشت است؟
پاسخ خود را بدون هیچ توضیحی ارائه دهید.

GROUND TRUTH	GPT-5 RESPONSE
تنبک ✓	دف ✗

Figure 5: An example from the PerCul-SAQ dataset (entry 1_O_MUS, Music category). The story implicitly encodes a Persian cultural artifact; GPT-5 incorrectly identifies دف (Daf) as the answer.

ID: **73** **Label: Normal** **Scope:** Specific

ENVIRONMENT

بازار → Bazaar

ORIGINAL DATASET — NORM & CONTEXT

FA: چانه‌زنی و مذاکره بر سر قیمت‌ها با فروشندگان

EN: Bargaining and negotiating prices with vendors

↓ We rewrote each original norm into a Short Answer Question (SAQ).

REWRITTEN SAQ PROMPT

در بازار ایران هنگام خرید با فروشندگان چه رفتاری رایج است؟

GROUND TRUTH	GPT-5 RESPONSE
چانه‌زنی ✓	چانه‌زنی بر سر قیمت ✓

Figure 6: An example from the ISN-SAQ dataset (entry 73). The original norm is shown as-is from the dataset; we rewrote it into a Short Answer Question (SAQ) for evaluation. The English translation is: "What behavior is common when shopping with vendors in Iranian bazaars?" GPT-5 correctly identifies چانه‌زنی (bargaining) as the expected social norm.