

ParsBench: A Benchmark for Evaluating Fine-Grained Culturally Grounded Commonsense Reasoning in Persian

Mehrnoush Alizade
Missouri S&T
malizade@mst.edu

Farbod Davoodi
Missouri S&T
fd4b9@mst.edu

Suman Kalyan Maity
Missouri S&T
smaity@mst.edu

Abstract

Large Language Models (LLMs) have achieved remarkable progress in natural language understanding and generation, yet their evaluation remains heavily centered on English and Western cultural contexts, limiting their reliability in culturally diverse and low-resource settings. We introduce *ParsBench*, a comprehensive benchmark for evaluating fine-grained culturally grounded commonsense reasoning in Persian. ParsBench consists of three complementary components: Expert-QA, Multimodal-QA, and Long-Form QA, collectively covering structured reasoning, multimodal cultural understanding, and open-ended generation across diverse cultural categories. We evaluate a broad range of multilingual and multimodal LLMs and observe substantial variation across models and cultural categories. Our results reveal persistent weaknesses in culturally grounded non-English reasoning and significant disparities between textual and multimodal understanding. Overall, the findings demonstrate that existing English-centric evaluations fail to capture important dimensions of cultural competence and highlight the need for culturally grounded benchmarks for more inclusive evaluation. Our code and dataset are available at <https://github.com/ParsBench-CCS/ParsBench>

1 Introduction

Large Language Models (LLMs) have rapidly become a transformative technology with applications spanning education, healthcare, research, industry, and everyday digital interaction. Despite their broad adoption, a major challenge remains the linguistic imbalance in the data used to train these models. Because textual resources are heavily concentrated in a small number of high-resource languages, particularly English, LLMs often encode knowledge and reasoning patterns

that are disproportionately aligned with dominant linguistic and cultural contexts. Prior studies have shown that such imbalance leads to substantial performance disparities in low-resource languages, where models frequently exhibit weaker reasoning capabilities, reduced factual reliability, and culturally misaligned behavior (Vayani et al., 2025). These limitations highlight the need for linguistically and culturally grounded resources tailored to underrepresented languages such as Persian.

Commonsense reasoning is a fundamental component of natural language understanding, enabling models to interpret implicit knowledge about everyday situations, social norms, intentions, causal relations, and human behavior. Although recent LLMs demonstrate strong general reasoning abilities, they continue to struggle when commonsense reasoning depends on culturally specific assumptions and lived experiences. This challenge is particularly pronounced in low-resource and non-Western settings, where both training data and evaluation benchmarks remain limited. As LLMs are increasingly deployed in socially sensitive and high-stakes applications, the lack of culturally grounded evaluation resources raises important concerns regarding reliability, inclusiveness, and cultural alignment.

Recent years have seen growing interest in Persian language models and evaluation benchmarks (Abbasi et al., 2023; Abaskohi et al., 2024; Rezaeimanesh et al., 2025). However, existing evaluations of Persian commonsense reasoning still rely heavily on translated datasets or narrowly scoped tasks that do not adequately capture culturally grounded reasoning. Direct translation of English-centric benchmarks often fails to preserve implicit cultural assumptions, social practices, and context-dependent interpretations that are essential for commonsense understanding in Persian. Moreover, most existing resources focus on isolated evaluation settings, such as multiple-choice

reasoning, social norms, or educational tasks, limiting their ability to comprehensively assess cultural reasoning across diverse domains and modalities.

To address this gap, we introduce **ParsBench**, a comprehensive benchmark for evaluating culturally grounded commonsense reasoning in Persian. ParsBench combines expert-authored question answering, long-form open-ended generation, and multimodal cultural reasoning within a unified evaluation framework. The benchmark includes multiple-choice, completion-based, open-ended, and multimodal tasks designed to evaluate how language models reason about culturally situated knowledge and visual concepts. All benchmark instances are written natively in Persian and curated through collaboration with native speakers to ensure cultural authenticity, linguistic naturalness, and regional diversity.

ParsBench covers 20 cultural domains and diverse fine-grained subtopics spanning everyday life, traditions, religion, history, geography, food, media, and cultural artifacts in Iranian society. Beyond textual reasoning, the benchmark also evaluates multimodal cultural understanding through text guided image selection and image-to-description tasks. To ensure quality and cultural authenticity, all benchmark instances undergo expert review and multi-stage validation.

We evaluate a diverse set of LLMs and MLLMs spanning multiple architectures and scales, including Aya, Gemma-3, Llama-3.3, GPT-5-mini, Mistral, Qwen-2.5, Aya-Vision, Qwen3-VL, Ministral, and Pixtral. Through extensive experiments across structured, long-form, and multimodal settings, ParsBench provides a comprehensive framework for evaluating culturally grounded reasoning in language models.

2 Related Work

2.1 Commonsense Reasoning in English

Early research on commonsense reasoning primarily focused on linguistic and pragmatic inference, as exemplified by the Winograd Schema Challenge (Levesque et al., 2012) and Winogrande (Sakaguchi et al., 2021), which evaluate pronoun co-reference resolution within rich social and discourse contexts. Subsequent benchmarks broadened this focus to general commonsense knowledge, including physical commonsense reasoning, which assesses models understanding of real-

world properties, affordances, and interactions (Bisk et al., 2020; Zellers et al., 2019b), and social commonsense reasoning, which tests the interpretation of human emotions, intentions, and social norms (Sap et al., 2019, 2018). More recent work has expanded commonsense evaluation to additional cognitive dimensions, including numerical reasoning (Lin et al., 2020; Akhtar et al., 2023), temporal reasoning (Tan et al., 2023; Zhou et al., 2019), causal reasoning (Du et al., 2022; Talmor et al., 2019), and multimodal reasoning (Zellers et al., 2019a). Despite their breadth, these benchmarks are overwhelmingly developed in English and implicitly reflect Western cultural assumptions, thereby limiting their applicability and validity for evaluating commonsense reasoning in non-Western languages and cultures, including the Persian-speaking world.

2.2 Commonsense Reasoning in Persian

Recent years have witnessed growing efforts to develop Persian benchmarks for evaluating language understanding, commonsense reasoning, cultural knowledge, and multimodal capabilities. As summarized in Table 1, existing resources differ substantially in their construction methodology, reasoning format, cultural scope, and support for open-ended or multimodal evaluation. Early Persian benchmarks primarily focused on general natural language understanding rather than culturally grounded reasoning. *ParsiNLU* (Khashabi et al., 2021a) introduced a broad collection of Persian NLU tasks spanning question answering, paraphrasing, natural language inference, sentiment analysis, and machine translation, providing an important foundation for evaluating Persian language models. Similarly, *PQuAD* (Darvishi et al., 2023) and *FarsTail* (Amirkhani et al., 2026) focused on extractive question answering and natural language inference using Persian Wikipedia and exam-based datasets. While valuable, these benchmarks were not designed to evaluate culturally grounded commonsense reasoning. More recent work has shifted toward commonsense and culturally informed reasoning. *PerCoR* (Alikhani et al., 2025) introduced a large-scale Persian commonsense sentence completion benchmark using automatically generated multiple-choice questions, while *FarSense* (Zeinalipour et al., 2025) expanded evaluation to explanation, causal reasoning, and counterfactual reasoning tasks. However, *FarSense* intentionally minimizes dialectal

Dataset	Size	Data Source/Origin	Construction	Question-Type	Scope/Categories	MM	Cult.	Tax.	Open.
PerCoR (Alikhani et al., 2025)	106K	News, cultural, and other Persian web sources	Automatic generation and filtering	MCQ / sentence completion	Broad topical coverage; no explicit cultural-category taxonomy	-	-	-	-
FarsEval-PKBETS (Shamsfard et al., 2025)	4,000	Existing Persian datasets, translated English samples, and newly collected Iranian educational/local questions	Benchmark compilation and human evaluation	MCQ / SA / DA	Medicine, law, religion, Persian language, encyclopedic knowledge, social knowledge, ethics/bias, and text generation	-	✓	-	✓
ParsiNLU (Khashabi et al., 2021b)	14.5K+	Persian task data collected across multiple NLU tasks	Manual and automatic task construction	MCQ / Paraphrasing / NLI / QA / Sentiment / MT	General Persian NLU; no explicit cultural-category taxonomy	-	-	-	-
PQuAD (Darvishi et al., 2023)	80K	Persian Wikipedia articles	Crowdsourced annotation	SQuAD-style extractive QA	Broad Wikipedia topics; no explicit cultural-category taxonomy	-	-	-	-
FarsTail (Amirkhani et al., 2026)	10,367	3,539 multiple-choice questions from Iranian university exams	MCQ-to-NLI conversion and cleaning	NLI	Exam-based topics such as religion, history, and constitution; not designed as a cultural benchmark	-	-	-	-
MELAC (Farsi et al., 2025)	66K	New Persian evaluation datasets focused on Iranian language and culture	Original translated sets	MCQ / Generation	Iranian law, Persian grammar, Persian idioms, university entrance exams, and related cultural-linguistic topics	-	✓	-	✓
PerCQA (Jamali et al., 2022)	989 Q / 21,915 A	Persian community forum data, mainly user questions and answers from Ninisite	Crawled and annotated community QA	Answer selection / Rating	Community QA on parenting, pregnancy, and daily-life concerns; no explicit cultural-category taxonomy	-	-	-	-
MEENA (Ghahroodi et al., 2026)	~10.4K	Persian educational exam questions with associated visual content, English version produced by translation	Multimodal benchmark construction	Visual MCQ	Mathematics, natural science, social science, humanities, reasoning, art/literature, and other school subjects	✓	-	-	-
FarSense (Zeinalipour et al., 2025)	4,210 passages / 29,470 items	Farsi Wikipedia lead paragraphs	LLM-assisted generation and human validation	TF / MCQ / Explanation / Cause-Effect / Counterfactual / Knowledge Completion	Commonsense scenarios; wording designed to minimize dialectal, cultural, and religious bias rather than define cultural categories	-	-	-	✓
TARAZ (Iranmanesh et al., 2026)	1,627	BLEnD-Fa, PerCul-SAQ, and ISN-SAQ Persian cultural datasets	Cultural adaptation and evaluation	QA and Short Answer	Everyday cultural knowledge, story-based cultural reasoning, and Iranian social norms	-	✓	-	✓
TAAROFBench (Sadr et al., 2025)	450	Native-speaker validated role-play scenarios for Persian taarof	Manual scenario design and validation	Role-play / Generation	Taarof-specific social norms across 12 interaction topics such as payment, gifts, dining, and compliments	-	✓	-	✓
ParsBench (Ours)	4,547	Manually curated Persian cultural questions, visual options, and community-informed cultural knowledge	Manual and community-driven construction	MCQ / LFQ / Multimodal QA	Politics, games, traditions, idioms, media, science, herbal knowledge, literature, music, education, official days, religion, history, geography, festivals, souvenirs, and food	✓	✓	✓	✓

Table 1: Comparison of Persian benchmarks and ParsBench. MM: multimodal input; Cult.: culture-specific content; Tax.: cultural taxonomy; Open.: open-ended generation.

and cultural specificity rather than explicitly modeling cultural knowledge. In parallel, culturally oriented benchmarks such as *MELAC* (Farsi et al., 2025), *TARAZ* (Iranmanesh et al., 2026), and *TAAROFBench* (Sadr et al., 2025) began evaluating Iranian social norms, moral alignment, and culturally situated interaction patterns. These resources represent important progress toward culturally aware Persian evaluation, but they remain limited either to specific domains, social behaviors, or single evaluation formats. Multimodal evaluation for Persian has also recently emerged through benchmarks such as *MEENA* (Ghahroodi et al., 2026), which evaluates Persian MLLMs on educational and reasoning-oriented visual tasks involving mathematics, science, diagrams, charts, and Persian art and literature. Unlike *MEENA*, which primarily emphasizes curriculum-based educational reasoning, ParsBench focuses on culturally grounded multimodal understanding. Specifically, our benchmark evaluates whether MLLMs can recognize fine-grained Persian cultural concepts across diverse domains and maintain consistent performance across both text-to-image and

image-to-text reasoning settings.

Overall, existing Persian benchmarks typically focus on isolated aspects of evaluation, such as general NLU, social norms, educational reasoning, or single-format commonsense tasks. In contrast, ParsBench provides a unified benchmark that combines fine-grained cultural taxonomy, multimodal reasoning, structured commonsense inference, and open-ended generation within a single culturally grounded evaluation framework.

3 ParsBench Dataset

ParsBench is a culturally grounded commonsense reasoning benchmark consisting of three complementary components: Expert-QA, Multimodal QA, and Long-Form Question Answering (LFQA) (see Appendix D for question distribution). Together, these components evaluate diverse aspects of cultural reasoning in Persian, including expert-authored commonsense questions, culturally grounded multimodal understanding, and open-ended generation requiring deeper explanation and justification.

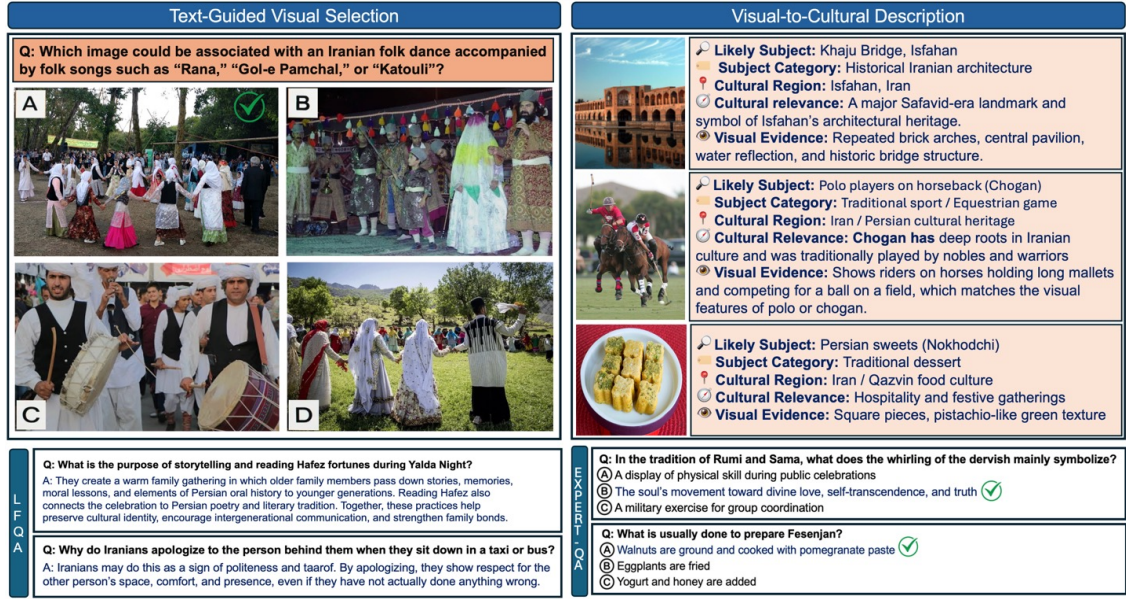


Figure 1: Examples from ParsBench across four task settings: (1) Text-Guided Visual Selection (TGVS), (2) Visual-to-Cultural Description (VCD), (3) Expert-style cultural reasoning (ExpertQA), and (4) Long-Form Question Answering (LFQA).

3.1 Expert-QA

The Expert-QA subset was constructed by native Persian speakers holding at least a graduate-level degree and with a minimum of ten years of formal education in Iran. Contributors were selected to ensure deep familiarity with Iranian cultural norms, social practices, and contextual knowledge relevant to the questions. This design choice ensures that the questions encode culturally grounded knowledge that cannot be recovered through surface-level translation or generic world knowledge alone. This dataset consists of 1,665 questions, systematically developed across 20 core cultural and conceptual categories (Fig 6).

3.1.1 Expert-QA Experimental Setup

Expert-QA questions are formulated in two complementary formats: direct multiple-choice questions (MCQ) and sentence completion. In the completion setting, the first sentence introduces a culturally grounded premise, and the second sentence must be completed in a manner that is logically consistent and semantically aligned with the given context. In the direct MCQ format, each item presents a clearly stated question with three candidate options, exactly one of which is correct. Models are required to select a single option, with no partial credit or multi-option selection permitted. To ensure quality and correctness, every question was independently reviewed by an additional do-

LLM	Overall		MCQ		Completion	
	En	Fa	En	Fa	En	Fa
Aya-8B	0.704	0.570	0.702	0.570	0.714	0.568
Aya-32B	0.858	0.854	0.861	0.849	0.847	0.878
Llama3.3-70B	0.858	0.864	0.853	0.860	0.885	0.885
Gemma3-12B	0.837	0.830	0.829	0.816	0.871	0.899
Gemma3-27B	0.875	0.854	0.872	0.844	0.892	0.902
Mistral-7B	0.723	0.745	0.721	0.750	0.728	0.718
Mistral-24B	0.772	0.801	0.764	0.800	0.812	0.805
Qwen2.5-7B	0.756	0.698	0.740	0.679	0.829	0.791
Qwen2.5-14B	0.699	0.673	0.681	0.652	0.787	0.774
Qwen2.5-72B	0.841	0.834	0.835	0.825	0.875	0.875
GPT-5 mini	0.878	0.862	0.872	0.856	0.909	0.891

Table 2: Accuracy of different LLMs on Expert-QA for Overall, MCQ, and Completion tasks under English (En) and Persian (Fa) prompts.

main expert.

3.1.2 Expert-QA LLM Performance

We analyze the LLMs’ performance across the 20 predefined cultural categories. Accuracy is first computed within each category and then averaged to obtain overall performance. This category-level analysis allows us to examine whether models exhibit systematic variation across different dimensions of culturally grounded knowledge rather than relying solely on aggregate scores.

Table 2 reports accuracy for each evaluated model under English (En) and Persian (Fa) prompts, further broken down by question format: overall, direct MCQ, and completion (C). Across models, performance on MCQ and completion formats remains broadly consistent with overall ac-

LLM	Overall		MCQ		Completion	
	En	Fa	En	Fa	En	Fa
Aya-8B-CoT	0.636	0.697	0.624	0.685	0.696	0.746
Aya-32B-CoT	0.856	0.852	0.851	0.847	0.881	0.879
Llama3.3-70B-CoT	0.825	0.853	0.815	0.850	0.874	0.867
Gemma3-12B-CoT	0.821	0.805	0.809	0.792	0.881	0.871
Gemma3-27B-CoT	0.856	0.844	0.848	0.835	0.895	0.888
Mistral-7B-CoT	0.530	0.675	0.521	0.662	0.574	0.753
Mistral-24B-CoT	0.789	0.755	0.778	0.741	0.843	0.825
Qwen2.5-7B-CoT	0.642	0.676	0.621	0.660	0.742	0.750
Qwen2.5-14B-CoT	0.718	0.666	0.696	0.651	0.718	0.737
Qwen2.5-72B-CoT	0.826	0.812	0.813	0.803	0.888	0.857
GPT-5 mini-CoT	0.875	0.870	0.868	0.862	0.909	0.909

Table 3: Accuracy of different LLMs on Expert-QA for Overall, MCQ, and Completion tasks under English (En) and Persian (Fa) CoT prompts.

curacy, suggesting that both settings capture comparably challenging forms of culturally grounded reasoning.

Several important trends emerge from the results. Larger models consistently outperform smaller ones, with top-performing systems such as GPT-5 mini, Llama3.3-70B, Gemma3-27B, and Aya-32B achieving overall accuracies above 0.85 in at least one language setting. GPT-5 mini achieves the strongest overall English performance (0.878) and the highest completion accuracy (0.909), while Llama3.3-70B achieves the best overall Persian performance (0.864) and the strongest Persian MCQ accuracy (0.860). These results indicate that strong general reasoning capabilities do not necessarily translate uniformly across languages and reasoning formats.

Figure 7(a) further reveals substantial variation across cultural categories, with certain domains showing consistently higher accuracy than others. Figure 7(b) highlights category-level differences between Persian and English prompts ($\Delta = \text{Acc}(\text{FA}) - \text{Acc}(\text{EN})$), demonstrating that prompt language effects are highly category-dependent. While models outperform their English-prompted counterparts in a small number of categories when evaluated in Persian, overall performance declines for most categories under Persian prompting, indicating persistent weaknesses in following non-English instructions and reasoning over culturally grounded content.

Table 3 presents Expert-QA performance under Chain-of-Thought (CoT) prompting. CoT prompting benefits several models, mostly in Persian completion-based reasoning, suggesting that explicit step-by-step reasoning can improve culturally grounded inference in low-resource settings. However, the gains are highly model-dependent, indicating that stronger reasoning prompts alone

Model	Base Acc.	QLoRA Acc.	Δ Acc.
Gemma-3-27B-it	86.28	88.45	+2.17
Gemma-3-12B-it	80.32	83.21	+2.89
Qwen2.5-7B-it	58.48	71.48	+13.00
Mistral-7B-it-v0.3	34.12	37.55	+3.43

Table 4: Base vs. QLoRA-adapted performance on the held-out ExpertQA test set.

are insufficient without robust multilingual alignment and cultural understanding.

3.1.3 Fine-tuning Results on ExpertQA

To examine the impact of task-specific adaptation, we fine-tuned Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, Gemma-3-12B-it, and Gemma-3-27B-it on the ExpertQA training split using QLoRA with a category-stratified 2:1 train-test split. Table 4 shows that QLoRA fine-tuning consistently improves performance across all models, with the largest gain observed for Qwen2.5-7B-Instruct, which improves from 58.48% to 71.48% (+13.00 points). Gemma-3-12B-it and Gemma-3-27B-it also achieve consistent gains (+2.89 and +2.17 points, respectively), while Mistral-7B-Instruct-v0.3 shows a smaller improvement (+3.43 points). Overall, these results demonstrate that parameter-efficient fine-tuning substantially improves culturally grounded commonsense reasoning in Persian, particularly for smaller and mid-sized models.

3.1.4 Error Analysis

To better understand model failure modes beyond aggregate performance, we conducted error analysis on ExpertQA. We categorized model failures into ten interpretable error types: i) errors in entity identification and attribution (confusing people, works, or titles), ii) geographic and regional association error (incorrectly linking places, foods, or customs to regions), iii) historical, temporal, and institutional reasoning error (errors involving dates, chronology, laws, and official facts), iv) cultural object composition and classification error (misidentifying ingredients, artifacts, or cultural categories), v) rituals and social custom error (festivals, ceremonies, etiquette, and traditions), vi) semantic and linguistic interpretation (idioms, terminology, and translations), vii) specialized cultural knowledge error (games, folk medicine, and traditional practices), viii) prototype-based guessing (choosing a famous or plausible answer instead of

Model	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
Aya-8B	7.32	11.98	10.15	8.65	9.65	5.82	2.83	0.67	0.17	42.76
Aya-32B	18.57	17.72	20.68	21.52	6.75	5.49	3.80	1.69	3.38	0.42
Qwen2.5-7B	16.24	19.63	16.92	17.09	7.45	5.41	6.43	0.34	1.52	8.97
Qwen2.5-14B	20.43	20.22	16.77	20.86	6.45	5.59	5.81	0.86	1.72	1.29
Qwen2.5-72B	14.33	19.22	16.94	17.92	7.82	7.82	9.12	1.63	3.26	1.95
Mistral-7B	12.10	20.46	18.79	22.01	10.55	5.41	6.69	0.26	0.39	3.35
Mistral-24B	17.20	20.37	15.87	16.67	9.52	5.56	7.41	1.06	1.32	5.03
Gemma-3-12B	14.58	23.05	20.00	16.95	8.14	4.75	6.78	0.34	2.71	2.71
Gemma-3-27B	18.57	21.94	16.88	15.61	9.28	4.22	4.22	0.00	5.06	4.22
GPT-5-mini	15.89	19.63	23.36	14.49	8.88	5.14	6.07	1.40	5.14	0.00

Table 5: Distribution (%) of error types across models on ExpertQA. E1–E10 denote the ten error categories defined in the error analysis.

the correct one), ix) partial-truth errors (providing related but incorrect facts), and x) reasoning/output failures (contradictory reasoning, rationale-answer mismatch, or invalid outputs).

Table 5 reveal several consistent patterns across models. Most failures arise from entity identification (E1), geographic and regional associations (E2), historical and temporal reasoning (E3), and cultural object classification (E4), indicating that models struggle primarily with retrieving and applying fine-grained culturally grounded knowledge rather than with general language understanding. In contrast, reasoning and output failures (E10) are relatively uncommon for most models, suggesting that modern LLMs generally produce coherent responses but often reason from incorrect cultural knowledge. An interesting exception is Aya-Expanse-8B, where E10 accounts for 42.76% of all errors, whereas GPT-5-mini exhibits virtually no reasoning/output failures (0.00%). We also observe that prototype-based guessing (E8) and partial-truth errors (E9) persist across nearly all models, highlighting a tendency to rely on plausible cultural associations instead of precise culturally grounded reasoning.

3.2 Multi-Modal QA Dataset

While textual benchmarks are important for evaluating cultural commonsense reasoning, many aspects of cultural knowledge are inherently visual and grounded in real-world artifacts, symbols, places, and practices. To comprehensively assess cultural understanding, models must therefore be able to connect textual reasoning with culturally meaningful visual evidence. To capture this dimension of cultural understanding, we introduce two complementary multimodal tasks: *Text-Guided Visual Selection (TGVS)* and *Visual-to-Cultural Description (VCD)*. These tasks assess whether models can connect culturally grounded textual and visual evidence while reducing re-

liance on superficial lexical cues or memorized textual patterns.

3.2.1 Experimental Setup

For each cultural category, we collected representative images from publicly available cultural resources¹. Images were organized by category to preserve semantic consistency and enable the construction of visually and culturally plausible distractors. This design makes the task substantially more challenging than simple category recognition, requiring models to identify fine-grained cultural concepts rather than rely on broad visual differences. The dataset covers a wide range of Persian cultural concepts, including traditional foods, festivals, musical instruments, landmarks, religion, literature, clothing, souvenirs, and social traditions.

The *Idioms* category posed a unique challenge because many Persian idioms are figurative and lack direct visual representations. To address this, we used the Gemini-3.1 Flash Image model to generate candidate images for selected idioms, followed by an expert-in-the-loop validation process involving native Persian speakers familiar with idiomatic usage. Annotators selected the image best representing each idiom, while contextual scenarios reflecting real-life usage were designed to ensure culturally appropriate interpretation.

For each category, we constructed two complementary multimodal tasks below.

Text-Guided Visual Selection (TGVS). In TGVS, models are given a culturally grounded textual description or scenario and must identify the correct image from visually plausible candidates. We further applied a controlled rewriting strategy inspired by Evol-Prompting and WizardLM (Xu et al., 2023) to improve clarity, reasoning difficulty, and distractor quality while preserving the original cultural intent. All rewritten questions were manually reviewed to ensure cultural relevance, interpretability, and visual answerability. This process encourages models to connect textual cues with fine-grained visual-cultural concepts rather than rely on superficial lexical or category-level matching.

Visual-to-Cultural Description (VCD). In VCD, models are given a culturally meaningful

¹We used category-specific searches from Wikimedia Commons to collect culturally relevant images for each dataset category.

image and must generate a structured textual description identifying the depicted cultural entity, object, place, or concept. Unlike TGVS, which maps text to images, VCD evaluates whether models can infer culturally grounded meaning directly from visual evidence.

3.2.2 Performance on Multimodal QA

To evaluate culturally grounded multimodal reasoning, we analyze the performance of the eight open-weight MLLMs spanning multiple architectures and scales: Aya-Vision-8B, Aya-Vision-32B (Dang et al., 2024), Gemma-3-12B-it, Gemma-3-27B-it, Qwen3-VL-8B-Instruct, Qwen3-VL-32B-Instruct (Yang et al., 2025), Ministral-3-14B-Instruct, and Pixtral-12B.

TGVS Task Performance. This evaluation aims to assess whether the MLLMs can effectively connect culturally grounded textual descriptions with visually meaningful Persian cultural concepts. We evaluate the Text-Guided Visual Selection (TGVS) task under both standard zero-shot prompting and Chain-of-Thought (CoT) prompting settings. Overall accuracies are summarized in Table 6, while category-level performance under Non-CoT and CoT prompting is reported in Tables 10 and 11, respectively.

Across both prompting settings, larger models generally achieve stronger performance, with Gemma-3-27B-it consistently obtaining the highest overall accuracy under zero-shot prompting (64.43%), followed closely by Gemma-3-12B-it (63.45%) and Qwen3-VL-32B-it (60.10%). Under CoT prompting, performance trends become more mixed: while some models such as Ministral-3-14B-it and Pixtral-12B show substantial gains (+4.83 and +21.28 points respectively), others such as Gemma-3-12B-it and Gemma-3-27B-it exhibit slight degradation.

The category-level results further reveal substantial variation across cultural domains. Certain visually grounded categories, such as *Festivals*, *Food*, *Religion*, and *Traditions*, achieve consistently high accuracy across strong models, likely because they contain distinctive visual-cultural cues. In contrast, categories such as *Politics*, *Science*, and *Idioms* remain considerably more challenging, indicating that culturally nuanced visual reasoning extends beyond simple object recognition and often requires deeper contextual understanding. The *Idioms* category is particularly diffi-

Models	Zero-shot Accuracy	CoT Accuracy
Aya-Vision-8B	28.87	30.05
Aya-Vision-32B	36.45	39.51
Gemma-3-12B-it	63.45	60.30
Gemma-3-27B-it	64.43	62.03
Ministral-3-14B-it	51.92	56.75
Qwen3-VL-32B-it	60.10	59.70
Qwen3-VL-8B-it	54.19	53.30
Pixtral-12B	25.91	47.19

Table 6: Overall accuracy (%) of MLLMs on TGVS task under zero-shot and CoT prompting.

Model	Exact Acc.(%)	VE-S=1(%)	VE-S=2(%)	VE-S=3(%)
Gemma-3-12B-it	29.35	8.15	30.90	60.94
Aya-Vision-32B	22.62	2.92	55.56	41.52
Gemma-3-27B-it	25.00	0.00	26.00	74.00
Ministral-3-14B-it	21.78	5.11	43.75	51.14
Qwen3-VL-32B-it	22.11	15.73	16.29	67.98
Aya-Vision-8B	10.39	7.23	45.78	46.99
Qwen3-VL-8B-it	8.70	8.57	7.14	84.29
Pixtral-12B	8.75	5.63	36.62	57.75

Table 7: Exact-match accuracy and Visual-Evidence Strength (VE-S) distribution for correct predictions in the VCD task.

cult across nearly all models, highlighting the challenge of interpreting figurative cultural concepts from visual representations alone. Together, these findings demonstrate that multimodal cultural reasoning remains highly uneven across both models and cultural domains.

VCD Task Performance. We further evaluate MLLMs on the VCD task, where models must infer culturally grounded entities directly from visual evidence and generate structured textual descriptions. Exact entity prediction remains substantially more challenging than TGVS, with Gemma-3-12B-it achieving the highest exact-match accuracy (29.3%), followed by Gemma-3-27B-it (25.0%), Aya-Vision-32B (22.6%), and Qwen3-VL-32B-it (22.1%) (Table 7). Smaller models such as Aya-Vision-8B (10.4%), Pixtral-12B (8.8%), and Qwen3-VL-8B-it (8.7%) struggle considerably.

Performance varies considerably across categories: visually distinctive domains such as *Traditions*, *Herbal*, and *Geography* achieve higher accuracy, while abstract categories such as *Idioms*, *Media*, and *Music* remain challenging (Table 12).

We additionally evaluate the **Visual-Evidence Strength (VE-S)** of correct predictions to measure whether generated explanations are visually grounded. Stronger models not only achieve higher exact-match accuracy but also produce stronger visual evidence. For example, Gemma-3-27B-it and Qwen3-VL-32B-it produce a high

proportion of strongly grounded explanations (VE-S=3), indicating better alignment between visual interpretation and textual reasoning. Under CoT prompting (Table 13), larger models further improve, particularly in culturally nuanced categories such as *Idioms*, *Music*, and *Souvenirs*, while smaller models show only marginal gains.

3.3 Long-form QA Dataset

To evaluate the cultural alignment of LLMs on open-ended generation tasks, we construct a Long-form Question Answering (LFQA) dataset using a semi-automatic pipeline. The objective of this dataset is to assess how effectively models generate extended, culturally grounded responses to questions related to Persian culture. We first identify Persian community-based websites relevant to each cultural category (Appendix C). These sources are selected to capture authentic, community-driven discourse rather than curated or encyclopedic content. After filtering candidate sources, we extract culturally relevant long-form questions together with their corresponding answers. By sourcing questions from naturally occurring community discussions and blogs, we ensure that the resulting questions reflect real information needs and culturally situated language use. Alongside each question, highly endorsed or widely accepted answers from the source websites are considered as reference responses and treated as ground-truth targets. All these collected QA pairs underwent an expert-in-the-loop quality control process. Two native Persian annotators jointly reviewed every instance for correctness, cultural authenticity, and linguistic quality. During this process, they rewrote and refined a subset of both the questions and reference answers to improve clarity, remove source-specific artifacts, normalize writing style, and better align the examples with the benchmark objectives while preserving their underlying meaning. As with the previous datasets, each cultural category in the LFQA benchmark contains paired question-answer instances, enabling consistent category-level analysis. The resulting dataset supports evaluation of both factual correctness and cultural coherence in long-form generation.

3.3.1 LFQA Experimental Setup

We evaluate multiple LLMs by prompting them to generate long-form responses for the LFQA questions. To assess answer quality beyond surface-

level correctness, we employ an LLM-as-judge framework that evaluates each response across multiple complementary dimensions. Specifically, generated answers are scored on informativeness, relevance, completeness, specificity, and correctness, capturing content coverage and factual alignment. We further evaluate coherence, clarity, conciseness, and depth of understanding to reflect structural quality and reasoning depth. For each evaluation dimension, we design detailed rubrics to guide the judgment process and ensure consistent scoring across models and questions (Appendix A.7). This multi-dimensional evaluation enables a fine-grained analysis of long-form generation quality that cannot be captured by accuracy-based metrics alone.

3.3.2 Performance of LLMs on LFQA

Table 8 summarizes LFQA performance across the nine evaluation dimensions using GPT-5.5 as the LLM-as-Judge. Overall, Gemma-3-12B achieves the highest overall score (3.83), followed by Gemma-3-27B and Aya-32B (both 3.75). Gemma-3-12B also achieves the highest scores in informativeness (4.10), relevance (4.24), and specificity (4.51), while Gemma-3-27B obtains the highest depth score (3.77). Aya-32B performs particularly strongly on coherence (4.18) and clarity (4.18).

GPT-5-mini achieves a competitive overall score of 3.45 and the third-highest correctness score (3.09), despite ranking below the strongest open-weight models on overall response quality. In contrast, smaller models generally show weaker performance, particularly in correctness, completeness, and depth. Mistral-7B records the lowest overall score (2.88) and correctness score (1.80), while Qwen2.5-7B also exhibits relatively low correctness (2.07). Across models, coherence and clarity are generally higher than completeness, correctness, and depth, suggesting that fluent and well-structured generation does not necessarily imply accurate or deeply reasoned culturally grounded responses. These results highlight the value of LFQA as a complementary evaluation setting that exposes differences in content quality and reasoning depth that may not be apparent from short-answer or multiple-choice accuracy alone.

Human-LLM Evaluation Agreement To assess the reliability of the LLM-as-Judge framework for LFQA evaluation, we compare LLM judge scores with human judgments on 250 ran-

LLM	Informative	Relevance	Completeness	Specificity	Correctness	Coherence	Clarity	Conciseness	Depth	Overall
Gemma-3-27B	3.90	4.11	3.23	3.76	3.46	4.17	4.15	3.22	3.77	3.75
Gemma-3-12B	4.10	4.24	3.39	4.51	2.99	4.07	4.31	3.26	3.62	3.83
Aya-32B	3.89	4.15	3.24	3.74	3.47	4.18	4.18	3.19	3.71	3.75
Aya-8B	3.56	3.61	2.80	3.63	2.36	3.80	4.02	3.31	2.88	3.33
Qwen2.5-72B	3.10	3.78	2.75	3.02	2.66	3.96	4.02	3.39	2.67	3.26
Qwen2.5-14B	3.43	3.82	2.80	3.79	2.21	3.90	4.02	3.74	2.84	3.40
Qwen2.5-7B	3.43	3.81	2.72	3.74	2.07	3.76	4.00	3.61	2.76	3.32
Mistral-24B	3.07	3.82	2.93	2.94	2.72	4.00	4.02	3.41	2.70	3.29
Mistral-7B	2.95	3.61	2.35	3.07	1.80	2.94	3.06	3.82	2.29	2.88
GPT-5-mini	3.35	4.01	2.83	3.46	3.09	3.99	4.02	3.31	2.99	3.45

Table 8: Performance of LLMs on LFQA using GPT-5.5 as the LLM-as-Judge. Responses are evaluated across informativeness, relevance, completeness, specificity, correctness, coherence, clarity, conciseness, depth of understanding, and overall quality.

Evaluator	Info.	Relevance	Completeness	Specificity	Correctness	Coherence	Clarity	Conciseness	Depth
Human 1 vs. LLM-Judge	0.988	0.955	0.951	0.933	0.950	0.900	0.802	0.965	0.978
Human 2 vs. LLM-Judge	0.824	0.866	0.902	0.937	0.891	0.890	0.917	0.936	0.880

Table 9: Per-metric correlation between human evaluators and GPT-5.5 as the LLM-as-Judge on 250 randomly selected LFQA responses.

domly selected instances, using the same evaluation rubric. Two human evaluators independently assessed all responses, achieving high inter-annotator agreement (Cohens $\kappa = 0.826$). As shown in Table 9, GPT-5.5 as LLM judge demonstrates consistently high per-metric agreement with human judgments, with correlations ranging from 0.802 to 0.988 across evaluation dimensions. The strongest agreement is observed for informativeness ($r = 0.988$), depth of understanding ($r = 0.978$), conciseness ($r = 0.965$), and relevance ($r = 0.955$), while even the lowest observed correlation, for clarity, remains substantial ($r = 0.802$). These results provide strong empirical support for the reliability of GPT-5.5 as an LLM-as-Judge for scalable evaluation of long-form, culturally grounded responses in ParsBench.

4 Conclusions

This work introduces ParsBench, which to the best of our knowledge, is the first multimodal benchmark that systematically targets fine-grained Persian cultural understanding. Unlike existing resources that primarily evaluate general recognition or broad cultural concepts, our dataset focuses on subtle and culturally specific cues across diverse Persian cultural domains, such as traditions, idioms, historical figures, literature, food, geography, festivals, official days, religion, and social practices. By emphasizing these nuanced elements, the dataset provides a new testbed for evaluating whether multimodal models can capture the depth, specificity, and contextual richness of Persian culture. Through three complemen-

tary components (ExpertQA, MultiModalQA, and Long-Form Question Answering), ParsBench provides a multi-dimensional evaluation framework that captures both answer correctness and the underlying reasoning processes across diverse cultural domains. Our extensive evaluation of state-of-the-art multilingual LLMs reveals substantial gaps in culturally grounded reasoning, significant performance disparities across models and cultural categories, highlighting persistent limitations that are obscured by English-centric or translation-based benchmarks. These findings underscore the necessity of native, culturally grounded evaluation resources for low-resource languages and demonstrate that robust cultural commonsense understanding remains an open challenge for current LLMs. We hope that ParsBench will serve as a foundational evaluation resource for advancing culturally inclusive AI research and for guiding the development of language technologies that better reflect the diversity of human knowledge and experience.

5 Limitations

While ParsBench provides a comprehensive framework for culturally grounded evaluation of LLMs, several limitations remain. First, the benchmark is centered on the social, cultural, and historical context of Iran, and therefore does not fully represent the diversity of the broader Persian-speaking world. Nevertheless, ParsBench includes selected examples related to Dari in Afghanistan and Tajikistan, the Rabatak inscription and the historical origins of Dari, and Persianate cultural and historical contexts such as Herat, Rudaki, Samarkand, Bukhara, and Greater Khorasan. It also contains a limited number of diaspora-related questions concerning Iranian cinema and music. These examples provide some representation beyond Iran but do not comprehensively capture the linguistic, cultural, and social diversity of Persian-speaking communities in Afghanistan, Tajikistan, and the diaspora. Expanding this coverage is an important direction for future work.

Second, portions of the benchmark, particularly TGVS and LFQA, involve limited LLM assistance during construction and LLM-based evaluation, which may introduce residual model bias. We mitigate this risk through multiple layers of human oversight. For TGVS, cultural categories and target concepts were defined by the authors, candidate images were retrieved based on relevant entities and visual cues, and all images were manually categorized and reviewed; LLMs were used only as supporting tools for retrieval and organization, with final inclusion decisions made by humans. For the more challenging idiom subset, 21 native Persian speakers evaluated candidate images, and only images receiving more than 70% agreement were retained. Despite these safeguards, we acknowledge that residual bias may still remain, particularly in LLM-assisted evaluation, and consider this an important limitation of the benchmark.

Third, because LFQA examples are derived from publicly available web sources, training-data contamination cannot be completely ruled out, particularly for proprietary LLMs with undisclosed pretraining data. We mitigate this risk by collecting examples from diverse Persian and English sources and applying expert-in-the-loop quality control. Two native Persian annotators jointly reviewed all QA pairs and rewrote a subset to improve clarity, remove source-specific artifacts, and preserve the intended meaning.

Finally, our multi-dimensional metrics and evaluation rubrics cannot capture every aspect of cultural competence, including subtle pragmatic cues, evolving social norms, and context-dependent interpretations. Future work should therefore extend ParsBench to broader Persian-speaking communities and interactive, longitudinal, and more context-sensitive evaluation settings.

References

- Amirhossein Abaskohi, Sara Baruni, Mostafa Masoudi, Nesa Abbasi, Mohammad Hadi Babalou, Ali Edalat, Sepehr Kamahi, Samin Mahdizadeh Sani, Nikoo Naghavian, Danial Namazifard, Pouya Sadeghi, and Yadollah Yaghoobzadeh. 2024. [Benchmarking large language models for Persian: A preliminary study focusing on ChatGPT](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia. ELRA and ICCL.
- Mohammad Amin Abbasi, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, and Behrouz Minaei Bidgoli. 2023. [Persianllama: Towards building first persian large language model](#). *Preprint*, arXiv:2312.15713.
- Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. 2023. [Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore. Association for Computational Linguistics.
- Morteza Alikhani, Mohammadtagha Bagherifard, Erfan Zivandi, and Mehran Sarmadi. 2025. [PerCoR: Evaluating commonsense reasoning in Persian via multiple-choice sentence completion](#). 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 2205–2224.
- Hossein Amirkhani, Mohammad AzariJafari, Soroush Faridan-Jahromi, Zeinab Kouhkan, Zohreh Pourjafari, and Azadeh Amirak. 2026. Farstail: A persian natural language inference dataset. *Soft Computing*, 30(2):891–903.
- Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. [Piqa: Reasoning about physical commonsense in natural language](#). *Proceedings of the AAAI Conference on Artificial Intelligence*.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and 1 others. 2024. Aya expanse:

- Combining research breakthroughs for a new multi-lingual frontier. *arXiv preprint arXiv:2412.04261*.
- Kasra Darvishi, Newsha Shahbodaghkhan, Zahra Abbasiantaeb, and Saeedeh Momtazi. 2023. Pquad: A persian question answering dataset. *Computer Speech & Language*, 80:101486.
- Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. [e-CARE: a new dataset for exploring explainable causal reasoning](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland. Association for Computational Linguistics.
- Farhan Farsi, Farnaz Aghababalo, Shahriar Shariati Motlagh, Parsa Ghofrani, MohammadAli Sadraei-Javaheri, Shayan Bali, Amir Hossein Shabani, Farbod Bijary, Ghazal Zamaninejad, AmirMohammad Salehoof, and 1 others. 2025. Melac: Massive evaluation of large language models with alignment of culture in persian language. 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 1933–1950.
- Omid Ghahroodi, Arshia Hemmat, Marzia Nouri, Seyed Mohammad Hadi Hosseini, Doratossadat Dastgheib, Mohammad Vali Sanian, Alireza Sahebi, Reihaneh Zohrabi, Mohammad Hossein Rohban, Ehsaneddin Asgari, and 1 others. 2026. Meena (persianmmmu): Multimodal-multilingual educational exams for n-level assessment. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 6457–6491.
- Reihaneh Iranmanesh, Saeedeh Davoudi, Pasha Abrishamchian, Ophir Frieder, and Nazli Goharian. 2026. Taraz: Persian short-answer question benchmark for cultural evaluation of language models. *arXiv preprint arXiv:2602.22827*.
- Naghme Jamali, Yadollah Yaghoobzadeh, and Heshaam Faili. 2022. [PerCQA: Persian community question answering dataset](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6083–6092, Marseille, France. European Language Resources Association.
- Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, Mozhdeh Gheini, Arman Kabiri, Rabeeh Karimi Mahabagdi, Omid Memarrast, Ahmadreza Mosallanezhad, Erfan Noury, Shahab Raji, Mohammad Sadegh Rasooli, Sepideh Sadeghi, and 6 others. 2021a. [Parsinlu: A suite of language understanding challenges for persian](#). *Transactions of the Association for Computational Linguistics*, 9:1147–1162.
- Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, Malihe Alikhani, Moin Aminnaseri, Marzieh Bitaab, Faeze Brahman, Sarik Ghazarian, and 1 others. 2021b. Parsinlu: a suite of language understanding challenges for persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.
- Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*. AAAI Press.
- Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. [Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.
- 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)*, Albuquerque, New Mexico. Association for Computational Linguistics.
- Nikta Gohari Sadr, Sahar Heidariasl, Karine Megerdoo-mian, 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.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. [Winogrande: an adversarial winograd schema challenge at scale](#). *Commun. ACM*.
- Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2018. [Atomic: An atlas of machine commonsense for if-then reasoning](#). *arXiv preprint*.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. [Social IQa: Commonsense reasoning about social interactions](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China. Association for Computational Linguistics.
- Mehrnoush Shamsfard, Zahra Saaberi, Seyed Mohammad Hossein Hashemi, Zahra Vatankhah, Mota-hareh Ramezani, Niki Pourazin, Tara Zare, Maryam Azimi, Sarina Chitsaz, Sama Khoraminejad, and 1 others. 2025. Farseval-pkbets: A new diverse benchmark for evaluating persian large language models. *arXiv preprint arXiv:2504.14690*.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense

- knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4149–4161. Association for Computational Linguistics.
- Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. [Towards benchmarking and improving the temporal reasoning capability of large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.
- Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadgign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, and 1 others. 2025. All languages matter: Evaluating llms on culturally diverse 100 languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19565–19575.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Kamyar Zeinalipour, Neda Jamshidi, Seyedehbahareh Hejazi, Marco Maggini, Monica Bianchini, Simone Paoletti, and Marco Gori. 2025. Farsense: A comprehensive commonsense benchmark and evaluation framework for the farsi language. 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 3529–3599.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019a. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6720–6730.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019b. [Hellaswag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.
- Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. “Going on a vacation” takes longer than “Going for a walk”: A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3361–3367. Association for Computational Linguistics.

A Additional Experiments, Prompts and Examples

A.1 Impact of Model Size on Accuracy and Correctness

To examine whether larger models consistently yield better performance, we analyze the relationship between model size and accuracy across all evaluation tasks in ParsBench. The results reveal a clear positive correlation between model scale and performance: larger models generally achieve higher accuracy and more reliable correctness across datasets. Figures 2, 3, 4 and 5 illustrate this trend, showing how model accuracy improves with increasing parameter size for Expert-QA, LFQA, as well as how correctness and response quality improve for long-form generation in the LFQA benchmark. These findings highlight the importance of model capacity for robust culturally grounded reasoning and long-form response generation.

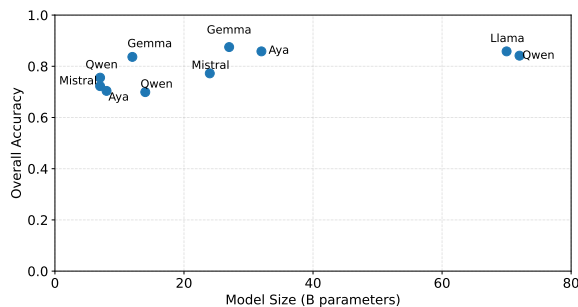


Figure 2: Overall accuracy plotted against model size (B parameters) On Expert-QA Dataset. Each point corresponds to a single model

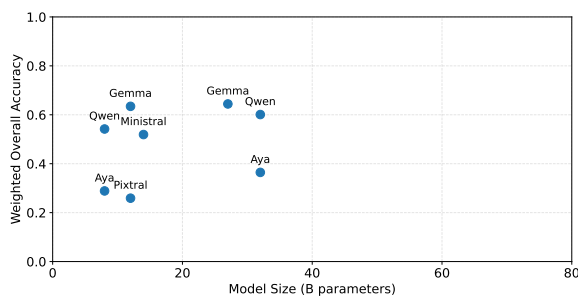


Figure 3: Overall accuracy plotted against model size (B parameters) on TGVS Dataset

A.2 Category-level A breakdown for Expert-QA

A detailed breakdown of the categories is presented in Figure 7.

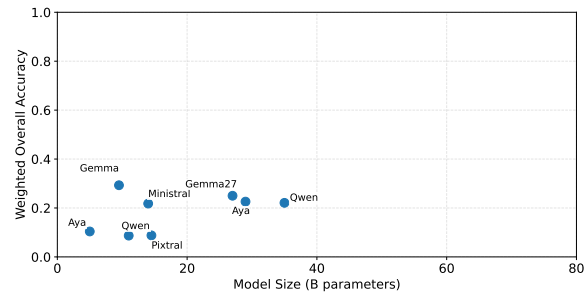


Figure 4: Overall accuracy plotted against model size (B parameters) on VCD Dataset

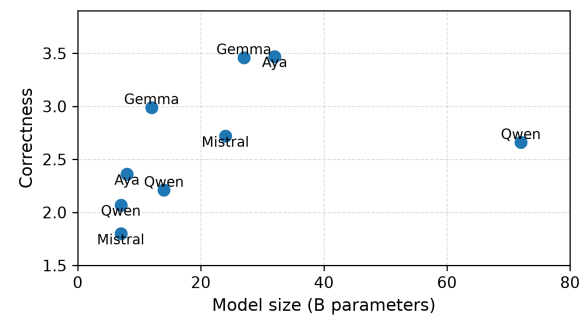


Figure 5: Mean Correctness scores (LLM-as-judge) as a function of model size on LFQA Dataset.

A.3 Expert-QA performance under Chain-of-Thought (CoT) prompting

Table 3 presents Expert-QA performance under Chain-of-Thought (CoT) prompting. Overall, CoT prompting produces mixed effects across models and languages. GPT-5 mini achieves the strongest performance across all evaluation settings, obtaining the highest overall accuracy in both English (0.875) and Persian (0.870), as well as the best MCQ and completion accuracies. Larger models such as Aya-32B, Llama3.3-70B, and Gemma3-27B also maintain strong and relatively stable performance under CoT prompting, suggesting that explicit reasoning prompts help preserve robust culturally grounded reasoning capabilities across



Figure 6: Main cultural components contributing to the identity and heritage of a country.

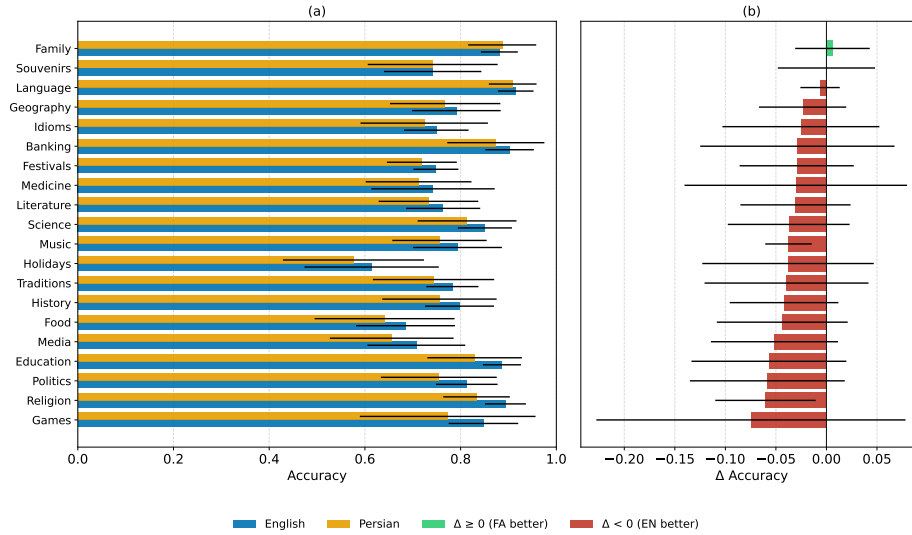


Figure 7: Category-level accuracy breakdown for Expert-QA across 20 cultural categories under English (En) and Persian (Fa) prompts. (a) Per-category accuracy for English and Persian. (b) Accuracy difference between Persian and English prompts ($\Delta = \text{Acc}(\text{Fa}) - \text{Acc}(\text{EN})$). Positive values indicate higher performance in Persian, while negative values indicate higher performance in English.

languages.

Interestingly, CoT prompting substantially benefits several smaller and mid-sized models in Persian. Aya-8B, Mistral-7B, and Qwen2.5-7B show notable improvements under Persian prompts compared to their English counterparts, particularly in the completion setting. For example, Mistral-7B improves from 0.530 (En) to 0.675 (Fa) overall accuracy, with a large gain in Persian completion accuracy (0.753). This suggests that explicit step-by-step reasoning may help smaller multilingual models better utilize culturally grounded contextual cues in Persian. However, CoT prompting does not uniformly improve performance for all models, as some larger systems such as Mistral-24B and Qwen2.5-72B exhibit slight degradation compared to their standard prompting results. Overall, the findings indicate that the effectiveness of CoT prompting depends strongly on model capacity, multilingual alignment, and reasoning robustness.

A.4 Fine-Tuning Results on ExpertQA

To investigate whether task-specific adaptation improves culturally grounded reasoning, we fine-tuned four open-weight LLMs - Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, Gemma-3-12B-it, and Gemma-3-27B-it on the ExpertQA training split using QLoRA. The dataset was divided using a category-stratified 2:1 train-test split to ensure balanced representation of all cultural categories

in both training and evaluation. To reduce positional bias, answer options were randomly shuffled prior to splitting, and models were trained to generate only one of the valid Persian answer keys.

Table 4 compares the performance of the base and QLoRA-adapted models on the held-out ExpertQA test set. Overall, QLoRA fine-tuning consistently improves performance across all evaluated models, although the magnitude of improvement varies substantially by architecture and model size. The largest gain is observed for Qwen2.5-7B-Instruct, which improves from 58.48% to 71.48% accuracy (+13.00 points), suggesting that smaller models benefit considerably from task-specific cultural adaptation. Gemma-3-12B-it and Gemma-3-27B-it also exhibit consistent gains (+2.89 and +2.17 points, respectively), indicating that even stronger base models retain room for improvement through supervised adaptation on culturally grounded reasoning tasks. In contrast, Mistral-7B-Instruct-v0.3 achieves a smaller improvement (+3.43 points), though fine-tuning substantially improves instruction-following stability by eliminating invalid outputs during evaluation.

These findings suggest that lightweight parameter-efficient fine-tuning can significantly improve both reasoning accuracy and adherence to constrained answer formats in Persian cultural commonsense reasoning tasks, particularly for smaller and mid-sized models.

Training Hyperparameters.

- Number of epochs: 3
- Learning rate: 2×10^{-4}
- Per-device training batch size: 2
- Gradient accumulation steps: 8
- Maximum sequence length: 1024 tokens
- LoRA rank (r): 16
- LoRA alpha: 32
- LoRA dropout: 0.05
- Optimizer: AdamW 8-bit (paged)
- Learning-rate scheduler: cosine decay
- Mixed precision: bfloat16 when available, otherwise float16

A.5 LLM Performance Across Cultural Categories on LFQA

To better understand how LLMs handle culturally grounded long-form generation across different knowledge domains, we analyze model performance at the level of individual cultural categories. This category-level evaluation provides deeper insight into the strengths and weaknesses of each model that are not visible from aggregate scores alone.

Figure 8 presents the resulting correctness scores for all evaluated models across cultural categories, offering a fine-grained comparative view of LLM performance on the LFQA benchmark. The results reveal substantial variation both across categories and across models, indicating that long-form cultural reasoning ability is highly non-uniform. While certain cultural domains consistently elicit strong responses from multiple models, others remain significantly more challenging, and these patterns vary markedly by model architecture and scale. This analysis underscores the importance of category-level evaluation for diagnosing culturally grounded reasoning behavior in long-form generation.

A.6 Multi-Modal Question Answering Prompts and Additional Results

A.6.1 Visual-to-Cultural Description Prompt

We use the following prompt to guide generation.

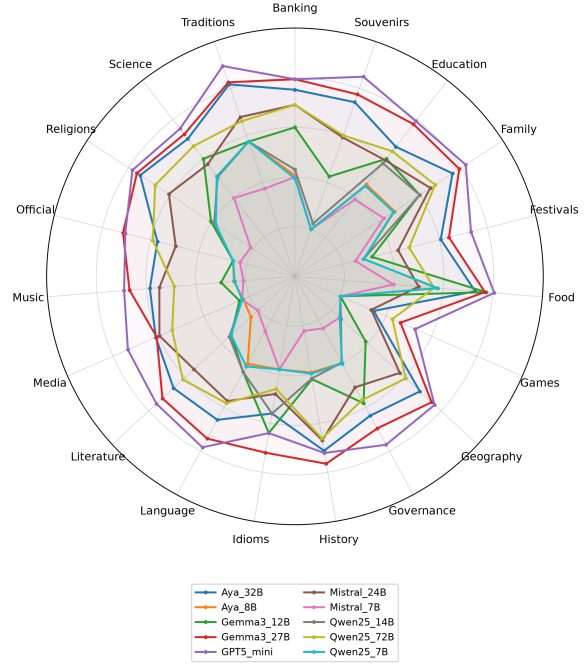


Figure 8: LLM-as-judge overall comparison across LLMs for various cultural categories.

Visual-to-Cultural Description Prompt

Analyze this image within a Persian cultural context. Identify the most likely primary subject shown in the image based only on the visible evidence. The subject may be a person, place, food, clothing, instrument, symbol, event, scene, or other culturally relevant visual element.

Write the output exactly in the following format and do not include any extra text:

Likely subject: $\langle \text{subject_name} \rangle$
 Subject category: $\langle \text{subject_category} \rangle$
 Visual evidence: $\langle \text{visual_evidence} \rangle$
 Confidence: $\langle \text{high/medium/low} \rangle$

If you do not know the exact subject/category, write “unknown” in the likely subject and subject category fields.

This structured output format reduces uncontrolled generation and facilitates more consistent comparison across models.

We evaluate the VCD task under a Chain-of-Thought (CoT) prompting setting to examine whether explicit reasoning improves culturally grounded visual interpretation. The results in Table 13 show that CoT prompting benefits several larger models, particularly Gemma-3-27B-it and Qwen3-VL-32B, which achieve the highest overall exact-match accuracies of 33.7% and 28.4%, respectively. Compared to the standard prompting setting, Gemma-3-27B-it exhibits a substantial im-

Category	Aya-Vision-8B	Aya-Vision-32B	Gemma3-12B	Gemma3-27B	Ministral3-14B	Qwen3-VL-32B	Qwen3-VL-8B	Pixtral-12B
Politics	18.1	22.7	39.9	43.9	34.8	34.8	36.3	22.7
Game	26.1	28.7	69.0	66.7	45.2	54.7	45.2	30.9
Traditions	39.1	43.2	74.3	74.3	74.3	68.9	63.5	25.7
Idioms	33.8	29.6	40.9	64.8	36.6	52.1	49.3	21.1
Media	21.1	32.8	55.7	55.7	37.7	50.8	55.7	21.3
Science	20.4	26.5	42.9	53.1	46.9	44.9	49.0	26.5
Herbal	19.4	35.8	70.1	74.6	50.7	68.7	59.7	31.3
Literature	33.3	28.7	62.1	66.7	59.1	63.6	59.1	18.2
Music	39.6	47.2	66.0	56.6	37.7	58.5	43.0	22.6
Education	23.1	32.7	53.9	55.8	40.4	38.5	42.3	34.6
OfficialDays	30.0	43.3	40.0	43.3	36.7	36.7	36.7	20.0
Religion	32.5	45.0	72.5	70.0	60.0	62.5	50.0	27.5
History	33.3	39.1	66.7	71.0	62.3	65.2	63.8	27.5
Geography	22.6	38.7	53.2	59.7	51.6	71.0	54.8	29.0
Festivals	39.1	56.5	97.8	97.8	82.6	95.7	82.6	17.4
Souvenirs	31.4	40.0	75.7	75.7	54.3	67.1	57.1	27.1
Food	27.8	37.1	83.5	73.2	59.8	70.1	57.7	32.0
Overall	28.87	36.45	63.45	64.43	51.92	60.10	54.19	25.91

Table 10: Category-wise accuracy (%) of MLLMs on Text-Guided Visual Selection (TGVS) task using standard zero-shot prompting.

Category	Aya-Vision-8B	Aya-Vision-32B	Gemma3-12B	Gemma3-27B	Ministral3-14B	Qwen3-VL-32B	Qwen3-VL-8B	Pixtral-12B
Politics	31.2	30.3	36.3	40.9	36.3	33.3	30.3	19.7
Game	33.3	52.8	57.1	59.5	59.5	52.8	54.7	45.2
Traditions	40.5	50.0	66.2	73.0	74.3	68.9	67.6	64.9
Idioms	38.0	43.7	39.4	45.1	47.9	49.3	50.7	42.3
Media	24.6	39.3	50.8	55.7	37.7	55.7	52.5	34.4
Science	14.3	24.5	36.7	49.0	42.9	46.9	46.9	42.9
Herbal	23.9	31.3	64.2	67.2	59.7	71.6	55.2	53.7
Literature	34.8	45.5	63.6	63.6	60.6	68.2	57.6	37.9
Music	32.1	47.2	54.7	58.5	52.8	67.9	32.1	35.9
Education	21.2	23.1	34.6	38.5	34.6	40.4	36.5	42.3
Official Days	23.3	36.7	36.7	50.0	33.3	33.3	40.0	46.7
Religion	47.5	45.0	67.5	70.0	62.5	55.0	52.5	52.5
History	36.2	44.9	60.9	69.6	69.6	62.3	59.4	53.6
Geography	21.0	24.2	51.6	54.8	58.1	66.1	54.8	38.7
Festivals	34.8	54.4	91.3	91.3	89.1	89.1	87.0	67.4
Souvenirs	34.3	38.6	61.4	62.9	62.9	64.3	57.1	60.0
Food	20.6	41.2	60.8	69.1	66.0	69.1	59.8	57.7
Overall	30.05	39.51	60.30	62.03	56.75	59.70	53.30	47.19

Table 11: Category-wise accuracy (%) of MLLMs on Text-Guided Visual Selection (TGVS) task using CoT prompting.

provement (+8.7 points), indicating that explicit reasoning prompts help the model better connect visual evidence with culturally grounded textual concepts.

The category-level results further reveal that CoT prompting is especially beneficial for categories requiring contextual or interpretive reasoning, such as *Idioms*, *Music*, *Souvenirs*, and *Food*. For example, Gemma-3-27B-it improves notably in *Idioms* and *Music*, suggesting that reasoning-based prompting helps models infer abstract or culturally nuanced concepts from visual cues. Similarly, Qwen3-VL-32B demonstrates strong performance in visually rich categories such as *Geography*, *Food*, and *Traditions*.

However, the improvements are not uniform across models. Smaller models such as Aya-Vision-8B and Qwen3-VL-8B show only marginal

gains, indicating that the effectiveness of CoT prompting depends strongly on model capacity and multimodal reasoning ability. Overall, these findings suggest that explicit reasoning prompts can enhance culturally grounded visual interpretation, particularly for larger MLLMs and culturally nuanced categories that require deeper contextual understanding.

A.7 LLM-as-Judge Evaluation Rubric for Long-form QA

LLM-as-a-Judge Evaluation Prompt

You are an expert evaluator with deep knowledge of {language} and its cultural context. Your task is to assess how well the system-generated answer addresses the question compared to the reference answer.

Question: {question}

Category	Aya-Vision-8B	Aya-Vision-32B	Gemma3-12B	Gemma3-27B	Ministral3-14B	Pixtral-12B	Qwen3-VL-8B	Qwen3-VL-32B
Politics	1	6	6	11	5	0	2	6
Game	6	18	2	10	27	3	0	12
Traditions	34	50	57	17	27	9	0	15
Idioms	0	0	3	1	0	0	0	0
Media	0	0	4	3	4	0	2	2
Science	6	12	28	19	4	3	15	22
Herbal	16	23	26	25	13	8	6	14
Literature	2	2	20	16	14	0	12	13
Music	0	2	0	2	0	0	1	9
Education	2	0	8	8	1	5	6	11
Official Days	3	5	12	5	8	2	1	8
Religion	2	7	18	14	14	6	1	11
History-	3	11	8	9	12	6	4	6
Geography	1	7	6	15	23	12	3	18
Festivals	2	5	7	10	6	2	2	6
Souvenirs	4	11	17	22	11	10	12	8
Food	1	12	11	13	7	5	3	17
Total Correct	83	171	233	200	176	71	70	178
Total Rows	799	756	794	800	808	811	805	805
Overall	10.4	22.6	29.3	25.0	21.8	8.8	8.7	22.1

Table 12: Category-wise exact entity match counts and weighted overall exact-match accuracy (%) of MLLMs on the Visual-to-Cultural Description (VCD) task.

Reference Answer: {reference _answer}

System Response: {system_response}

Rate the system response on the following metrics using a **1–5 scale**.

A. Informativeness (1–5) Assesses whether the answer provides meaningful and useful information.

- 1: Non-informative; irrelevant, empty, or missing information.
- 2: Low informativeness; minimal useful content, mostly generic.
- 3: Moderately informative; some useful information but lacks depth.
- 4: Informative; covers major points with sufficient detail.
- 5: Highly informative; rich, detailed, and adds meaningful depth.

B. Relevance (1–5) Measures how directly the answer addresses the question.

- 1: Off-topic; does not answer the question.
- 2: Low relevance; largely unrelated or tangential.
- 3: Moderately relevant; addresses the question but includes irrelevant content.
- 4: Relevant; mostly focused with minor drift.
- 5: Highly relevant; fully focused with no extraneous information.

C. Completeness (1–5) Evaluates whether all parts of the question are addressed.

- 1: Very incomplete; misses most key aspects.
- 2: Partially incomplete; addresses only a small portion.

- 3: Moderately complete; covers several aspects with noticeable gaps.

- 4: Mostly complete; addresses nearly all important parts.

- 5: Fully complete; thoroughly covers all components.

D. Specificity (1–5) Assesses whether the answer avoids vagueness and generality.

- 1: Very vague; generic or abstract.
- 2: Somewhat vague; limited specificity.
- 3: Moderately specific; generally specific with some general statements.
- 4: Specific; clear and precise with minor generalization.
- 5: Highly specific; fully tailored, concrete, and precise.

E. Correctness (1–5) Measures factual alignment with the reference answer.

- 1: Incorrect; contradicts or misrepresents the reference.
- 2: Partially incorrect; major factual errors.
- 3: Mostly correct; minor errors or missing evidence.
- 4: Correct; accurate with negligible issues.
- 5: Fully correct; completely accurate and consistent.

F. Coherence (1–5) Evaluates logical organization and structure.

- 1: Incoherent; no logical flow.

Category	Aya-8B	Aya-32B	Gemma-12B	Gemma-27B	Ministral-14B	Pixtral-12B	Qwen3-VL-8B	Qwen3-VL-32B
Politics	2	6	4	8	4	1	3	8
Game	5	15	9	14	17	8	1	17
Traditions	11	55	32	23	30	5	3	35
Idioms	2	5	8	10	1	6	0	6
Media	0	0	3	2	2	0	2	0
Science	4	8	20	23	14	4	17	20
Herbal	14	7	20	26	14	8	7	9
Literature	0	2	20	16	15	0	14	13
Music	2	7	13	19	3	7	2	17
Education	5	5	9	10	9	4	6	13
OfficialDays	4	5	11	7	9	4	1	8
Religion	0	6	12	15	13	3	3	9
History	3	8	7	11	10	8	3	8
Geography	4	5	7	13	12	7	2	20
Festivals	2	5	7	7	9	2	1	6
Souvenirs	5	11	15	32	13	4	2	17
Food	15	13	21	31	15	11	9	22
Total Correct	78	163	218	267	190	82	76	228
Total Rows	795	800	793	793	793	800	803	804
Overall	9.8	20.4	27.5	33.7	24.0	10.2	9.5	28.4

Table 13: Category-wise exact entity match counts and weighted overall exact-match accuracy (%) of MLLMs on the Visual-to-Cultural Description (VCD) task in CoT setting.

- 2: Poorly coherent; disorganized and difficult to follow.
- 3: Moderately coherent; some structure with logical gaps.
- 4: Coherent; clear flow with minor issues.
- 5: Highly coherent; seamless logical progression.

G. Clarity (1–5) Measures ease of understanding.

- 1: Unclear; confusing or incomprehensible.
- 2: Mostly unclear; ambiguous phrasing.
- 3: Somewhat clear; generally understandable with confusion.
- 4: Clear; easy to understand with minor issues.
- 5: Highly clear; exceptionally clear and precise.

H. Conciseness (1–5) Assesses efficiency of expression.

- 1: Not concise; excessively verbose.
- 2: Somewhat verbose; unnecessary repetition.
- 3: Moderately concise; reasonable length with some redundancy.
- 4: Concise; minimal redundancy.
- 5: Highly concise; optimal length with no wasted content.

I. Depth of Understanding (1–5) Evaluates conceptual and cultural understanding.

- 1: Superficial; misunderstands core concepts.
- 2: Shallow; basic understanding without nuance.
- 3: Adequate; correct but limited depth.
- 4: Deep; strong understanding with meaningful insights.
- 5: Highly deep; expert-level, nuanced understanding.

Output format (strict JSON only):

```
{
  "informativeness": <1-5>,
  "relevance": <1-5>,
  "completeness": <1-5>,
  "specificity": <1-5>,
  "correctness": <1-5>,
  "coherence": <1-5>,
  "clarity": <1-5>,
  "conciseness": <1-5>,
  "depth_understanding": <1-5>
}
```

A.8 Example of LFQA/ Expert-QA

Table 14 provides representative examples from the LFQA and Expert-QA subsets of our proposed Persian cultural QA benchmark. The examples are included to illustrate the linguistic and cultural structure of the benchmark, as well as the difference between the two task formats. In each example, the original Persian question or prompt is shown together with its English translation to make the benchmark accessible to both Persian-speaking and non-Persian-speaking readers.

The LFQA examples require models to generate long-form answers that explain culturally grounded practices, social norms, traditions, or meanings. These questions are designed to evaluate whether a model can produce coherent and culturally faithful explanations rather than only selecting a short answer. In contrast, the ExpertQA examples follow a multiple-choice format and evaluate culturally grounded reasoning under a constrained answer space. This setting allows us to test whether models can identify the correct culturally relevant option among plausible alternatives.

Together, these examples demonstrate the complementary nature of the two subsets: LFQA evaluates open-ended cultural explanation, while Expert-QA evaluates structured cultural reasoning and answer selection.

B Cross-Task Comparisons

To better understand whether model capabilities generalize across different reasoning and generation settings, we analyze the consistency of model rankings across multiple ParsBench tasks. Specifically, we examine whether models that perform well on structured reasoning benchmarks also maintain strong performance in open-ended generation and multimodal cultural understanding tasks. These analyses provide insight into the extent to which different tasks capture shared or complementary aspects of culturally grounded reasoning.

B.1 Cross-Task Comparison Between ExpertQA and LFQA

Although both ExpertQA and LFQA evaluate culturally grounded commonsense understanding, they differ substantially in output structure and evaluation criteria. ExpertQA focuses on constrained reasoning and answer correctness, whereas LFQA emphasizes open-ended generation, coherence, and culturally aligned explanation. Therefore, exact consistency in model rankings across the two tasks is not necessarily expected. Instead, our goal is to assess whether relative performance trends remain stable across structured and long-form reasoning settings.

To quantify ranking consistency, we convert model performance within each category into ordinal rankings and compute Kendalls τ and Spearman's ρ . As shown in Table 15, the two tasks exhibit generally positive rank correlations, with statistically significant agreement in 11 of 19 cate-

gories under Kendalls τ . The strongest alignment occurs for *Religions* ($\tau = 0.857$, $\rho = 0.929$), followed by *Education* and *Music* ($\tau = 0.786$, $\rho = 0.905$). Significant positive correlations are also observed for *Souvenirs*, *Games*, *Geography*, *History*, *Idioms*, *Literature*, *Language*, *Media*, and *Science*. In contrast, weaker or non-significant correlations appear in categories such as *Banking*, *Official Days*, and *Festivals*, indicating that strong performance on constrained reasoning tasks does not always transfer directly to open-ended cultural generation.

B.2 Cross-Task Comparison Between TGVS and VCD

We further analyze whether model rankings remain consistent across the two multimodal tasks: Text-Guided Visual Selection (TGVS) and Visual-to-Cultural Description (VCD). Although both tasks evaluate culturally grounded visual understanding, they differ fundamentally in task direction and reasoning requirements. TGVS evaluates whether models can map culturally grounded textual cues to appropriate images, whereas VCD evaluates whether models can infer culturally meaningful textual descriptions directly from visual evidence.

To assess ranking consistency, we convert model performance in each task into ordinal rankings and compute Kendalls τ and Spearman's ρ correlations across categories. The results are presented in Table 16. The results reveal moderate-to-strong rank agreement in several visually grounded categories. In particular, *Festivals* exhibits the strongest alignment ($\tau = 0.80$, $\rho = 0.88$), followed by *Herbal*, *Religion*, and *Souvenirs*, suggesting that models that successfully align cultural text with images also tend to generate stronger culturally grounded visual descriptions in these domains. However, categories such as *Games*, *History*, and *Traditions* exhibit weak or inconsistent correlations, indicating that success in text-to-image matching does not necessarily translate into accurate visual-to-text cultural interpretation. Overall, these findings suggest that TGVS and VCD capture complementary aspects of multimodal cultural reasoning and jointly provide a more comprehensive evaluation of cross-modal cultural understanding.

Question / Statement	Reference Answer / Options
LFQA: Long-form culturally grounded question answering	
جواب «زیارت قبول؟» چیست؟ What is an appropriate response to "Ziyarat qabool"?	قربان شما، ان شاء الله قسمت شما هم بشه. سلامت باشید، خدا شما را هم قسمت کنه. A common response is to thank the speaker and wish that they also have the opportunity to make the pilgrimage.
نان و پنیر و سبزی در سفره عقد نشانه چه چیزی است؟ What do bread, cheese, and herbs symbolize in the Iranian wedding spread?	نان و پنیر و سبزی در سفره عقد برای برکت زندگی استفاده می‌شود. نان و پنیر نشانه وسعت روزی هستند و سبزی نماد سرسبزی و سلامتی است. Bread, cheese, and herbs symbolize blessing, livelihood, prosperity, greenery, and health in the couple's shared life.
معنی ضرب‌المثل «آب که سر بالا بیره، قورباغه ابو عطا می‌خونه» چیست؟ What does the Persian proverb "If water flows uphill, the frog will sing Abu Ata" mean?	این ضرب‌المثل به غیرممکن بودن کاری اشاره دارد؛ زیرا آب برخلاف جهت سرایشی به سمت بالا حرکت نمی‌کند. این عبارت زمانی به کار می‌رود که بخواهند بگویند کاری با محاسبات و شرایط موجود پیش نخواهد رفت. The proverb refers to something impossible or unrealistic, especially when a plan cannot succeed under the given conditions.
راز برج‌های بادگیر، خنک‌کننده‌های جاودان کویر چیست؟ What is the secret behind Iranian windcatchers, the timeless cooling systems of the desert?	بادگیرها با شناخت رفتار باد، اختلاف دما و تغییر آب، جریان هوا را به درون بنا هدایت کرده و تهریه طبیعی ایجاد می‌کنند. Windcatchers use airflow, temperature differences, and evaporative cooling to create natural ventilation in desert architecture.
سازهای ایرانی در دوره ایران باستان چه نام داشتند و چگونه بودند؟ آیا هنوز این سازها نزد ایرانیان هستند؟ What were some musical instruments in ancient Iran, and are they still known among Iranians today?	از سازهای باستانی خوش‌آوا می‌توان از چنگ نام برد که همچنان پس از هزاران سال کاربرد دارد. ساز دیگری است که آواز آن مردم را به آسمان‌ها می‌برد. ساز دیگر بریت است که آن نیز آوازی دلنشین دارد. Ancient Iranian musical instruments include the chang, which has remained known for thousands of years, the ney, whose sound is often associated with a spiritual and elevated quality, and the barbat, another instrument known for its pleasant and expressive sound.
چرا افراد ایرانی وقتی در تاکسی یا حتی اتوبوس می‌نشینند، از پشت‌سری خود عذرخواهی می‌کنند؟ Why do Iranians apologize to the person behind them when they sit down in a taxi or even on a bus?	این رفتار نوعی تعارف و رعایت ادب اجتماعی در فرهنگ ایرانی است. وقتی فردی در وسیله نقلیه می‌نشیند، ممکن است تصور کند با حرکت خود مزاحم فضای شخصی یا دید فرد پشت‌سر شده است؛ بنابراین با گفتن «بخشید» یا عذرخواهی کوتاه، احترام و ملاحظه خود را نشان می‌دهد. This behavior reflects politeness and social consideration in Iranian culture. When someone sits down in a vehicle, they may feel that their movement has slightly disturbed the person behind them or blocked their space or view. A brief apology such as "excuse me" signals respect, courtesy, and awareness of others.
Expert-QA: Multiple-choice culturally grounded reasoning	
دره ستارگان یکی از پدیده‌های طبیعی و شگفت‌انگیز در جنوب ایران است. The Valley of Stars is one of the remarkable natural phenomena in southern Iran.	الف: دره ستارگان در جزیره قشم واقع شده است. ب: دره ستارگان در شمال ایران قرار دارد. ج: دره ستارگان در کوه‌های البرز ساخته شده است. A: The Valley of Stars is located on Qeshm Island. B: The Valley of Stars is located in northern Iran. C: The Valley of Stars was built in the Alborz Mountains.
نگار با دوستانش الک‌دولک بازی می‌کند. Negar is playing alak-dolak with her friends.	الف: نگار دولک را روی دو سنگ می‌گذارد و با چوب الک به آن ضربه می‌زند. ب: نگار توپ را درپیل می‌کند و به سمت دروازه شوت می‌زند. ج: نگار توپ پلاستیکی را به سمت حریفان پرت می‌کند تا آن‌ها را از بازی خارج کند. A: Negar places the dolak on two stones and hits it with the alak stick. B: Negar dribbles the ball and shoots toward the goal. C: Negar throws a plastic ball at opponents to eliminate them from the game.
معنی تکایه «سنگ مفت، گنجشک مفت» چیست؟ What does the Persian expression "free stone, free sparrow" mean?	الف: چون کاری بی‌هزینه است، بی‌محابا آن را امتحان می‌کنند. ب: تشویق به سخت‌کوشی و برنامه‌ریزی دقیق. ج: دعوت به شکار گنجشک با سنگ. A: Since the action costs nothing, people may try it without hesitation. B: Encouragement for hard work and careful planning. C: An invitation to hunt sparrows with stones.
به شهری سفر کردم که آرامگاه حافظ و سعدی در آن قرار دارد و به شهر گل و بلبل معروف است. I traveled to a city where the tombs of Hafez and Saadi are located, known as the city of flowers and nightingales.	الف: این شهر تبریز است. ب: این شهر شیراز است. ج: این شهر کرمان است. A: The city is Tabriz. B: The city is Shiraz. C: The city is Kerman.
نان محلی رایج و اصیل مردم بلوچستان چه نام دارد؟ What is the traditional and authentic local bread of the people of Balochistan called? MCQ	A: قشک نان / Qashk bread B: لوش نان / Lavash bread C: باگت نان / Baguette bread
امشب مراسم حنابندان دوست من است. Tonight is my friend's henna-bandān ceremony.	الف: حنابندان همیشه بعد از عروسی برگزار می‌شود. ب: جشن حنابندان ارتباطی با عروسی ندارد. ج: حنابندان معمولاً پیش از مراسم عروسی، به عنوان آیین پیش‌عروسی برگزار می‌شود. A: Henna-bandān is always held after the wedding. B: Henna-bandān is unrelated to weddings. C: Henna-bandān is usually held before the wedding ceremony as a pre-wedding ritual.

Table 14: Representative examples from the proposed Persian cultural QA benchmark. Each cell presents the original Persian text followed by its English translation. LFQA requires long-form cultural explanations, while Expert-QA uses multiple-choice cultural reasoning.

Category	Kendall's τ	p -value	Spearman ρ	p -value
Banking	0.429	0.179	0.643	0.086
Souvenirs	0.643	0.031	0.786	0.021
Education	0.786	0.006	0.905	0.002
Family	0.643	0.031	0.810	0.015
Festivals	0.429	0.179	0.643	0.086
Food	0.500	0.109	0.571	0.139
Games	0.643	0.031	0.786	0.021
Geography	0.643	0.031	0.810	0.015
Politics	0.500	0.109	0.667	0.071
History	0.643	0.031	0.786	0.021
Idioms	0.643	0.031	0.833	0.010
Language	0.500	0.109	0.714	0.047
Literature	0.643	0.031	0.786	0.021
Media	0.571	0.061	0.714	0.047
Music	0.786	0.006	0.905	0.002
Official days	0.357	0.275	0.262	0.531
Religions	0.857	0.002	0.929	0.001
Science	0.571	0.061	0.810	0.015
Traditions	0.500	0.109	0.738	0.037

Table 15: Rank correlation between ExpertQA and LFQA model rankings across categories.

Category	Kendall's τ	p -value	Spearman ρ	p -value
Politics	0.509	0.093	0.654	0.078
Game	-0.118	0.533	-0.179	0.670
Traditions	0.264	0.372	0.268	0.520
Idioms	0.369	0.249	0.452	0.260
Media	0.538	0.062	0.721	0.043
Science	0.254	0.328	0.371	0.365
Herbal	0.691	0.017	0.862	0.005
Literature	0.336	0.280	0.456	0.256
Music	0.536	0.081	0.523	0.182
Education	0.472	0.105	0.586	0.126
OfficialDay	0.320	0.297	0.434	0.281
Religion	0.618	0.034	0.790	0.019
History	0.254	0.328	0.251	0.547
Geography	0.285	0.398	0.428	0.289
Festivals	0.800	0.009	0.875	0.004
Souvenirs	0.592	0.044	0.698	0.053
Food	0.500	0.108	0.666	0.071

Table 16: Rank correlation between TGVS and VCD model rankings across categories.

C Forums and Web Communities Used for Data Collection

The Persian and English forums and web communities used to extract long-form questions include: Arq.ir, p30download.ir, daneshchi.ir, Boursieplus.ir, davoodvatani.com, vakilcity.app, 4barande.ir, delgarm.com, Stackexchange.com, Reddit.com, Quora.com, soorban.com, soja.ai, top-forum.ir, itfarzanegan.com, MehreMihan.ir, IranJwish.ir, my.pasokhgoo.ir, Itoll, ninisite.com, heyvalaw.com, and Abrangsafar.com. In addition, manual question extraction by Persian cultural experts relied on a set of curated web references, including sooremag.ir, atdcompany.ir, dokantick.com, fa.wikipedia.org, fa.wikishia.net, farapedia.ir, ghazaland.com, hamgardi.com,

Category	ExpertQA	LFQA	TGVS	VCD
Food and Drink	267	112	75	97
History	159	132	17	69
Literature	138	59	37	66
Idioms	100	10	54	71
Geography and Climate	99	31	76	62
Media and Entertainment	94	50	53	61
City Souvenirs and Local Products	80	24	51	70
Traditions and Customs	76	23	64	74
Music and Authentic Iranian Songs	71	73	25	53
Science and Technology	67	43	83	49
Religions	58	155	50	40
Games and Play Activities	58	56	34	42
Family Behaviors and Interpersonal Relations	54	43	–	–
Official Holidays and National Days	57	21	15	30
Governance and Society (Politics)	57	85	55	66
Traditional and Herbal Medicine	55	–	51	67
Language	54	63	–	–
Festivals and Celebrations	51	35	12	46
Education and Career	41	26	41	52
Banking	29	33	–	–
Total	1665	1074	793	1015

Table 17: Category-wise distribution of questions across ParsBench components. Categories with no instances in a particular component are marked with “–”

irajjamalimusic.com, irancook.com, iranmusicology.com, jobvision.ir, karboom.io, lastsecond.ir, loobli.co, matinabad.com, namnak.com, navakade.com, rayad.org, tarikHEMA.org, vi-giato.net, beytoote.com, digikala.com, eligasht.com, state.gov, and zoomg.ir.

D Category-wise Question Distribution

Table 17 presents the distribution of questions across the cultural categories and the four ParsBench components: ExpertQA, LFQA, TGVS, and VCD. The distribution reflects the different scopes of the components while maintaining broad coverage of culturally grounded topics. Overall, the benchmark covers a diverse set of cultural domains, including food, history, literature, religion, traditions, geography, media, music, and social practices.

E Annotator Details

The dataset was constructed by native Persian speakers holding at least a graduate-level degree and with a minimum of ten years of formal education in Iran (aged between 25-38). Contributors were selected to ensure deep familiarity with Iranian cultural norms, social practices, and contextual knowledge relevant to the questions. No payment was provided to the annotators.

For the ExpertQA component, five native Persian speakers independently developed culturally grounded questions and corresponding answers. Each question was subsequently reviewed by an additional expert to verify its correctness, cultural relevance, and clarity before inclusion in the

benchmark. For LFQA, questions and reference answers were collected from external Persian community sources. Two annotators performed quality control jointly by reviewing the collected question-answer pairs for accuracy, relevance, and linguistic quality. When necessary, they edited the questions and answers to improve clarity while preserving their original meaning. For the Multimodal QA component (TGVS and VCD), two annotators verified all question-image pairs to ensure that the visual content accurately represented the intended cultural concept and that the questions were culturally appropriate and unambiguous.

We did not compute a formal inter-annotator agreement as the annotation process was based on expert creation followed by expert validation rather than independent labeling of the same instances. Instead, disagreements were resolved through discussion until consensus was reached, and only consensus-approved instances were included in the final benchmark.

F AI assistance in Writing

AI Assistants were used solely to assist in improving the writing clarity and language of this paper. Specifically, AI-assisted refinements were applied to enhance readability, coherence, and grammatical accuracy. No AI-generated content was used to replace critical thinking or fabricate results. Ideas, methodology, experimental design, analysis, and conclusions were entirely conceived, developed, and executed by the authors.