Evaluating LLM Capabilities in Low-Resource Contexts: A Case Study of Persian Linguistic and Cultural Tasks

Jasmin Heierli, Rebecca Bahar Ganjineh, Elena Gavagnin
Zurich University of Applied Sciences, Winterthur, Switzerland
heej@zhaw.ch, ganjireb@students.zhaw.ch, gava@zhaw.ch

Abstract

We evaluate four representative large language models, namely GPT-4o, Gemini, Llama, and DeepSeek on a suite of linguistic and cultural tasks in Persian, covering grammar, paraphrasing, inference, translation, factual recall, analogical reasoning, and a Hofstede-based cultural probe under direct and role-based prompts. Our findings reveal consistent performance declines, alongside systematic misalignment with Iranian cultural norms. Role-based prompting yields modest improvements but does not fully restore cultural fidelity. We conclude that advancing truly multilingual models demands richer Persian resources, targeted adaptation, and evaluation frameworks that jointly assess fluency and cultural alignment.

1 Introduction

Despite rapid advances in large language models (LLMs), their multilingual capabilities remain deeply uneven. For dominant languages such as English, modern LLMs exhibit high performance in linguistic fluency, factual accuracy, and sociocultural alignment (Jin et al., 2024; Lai et al., 2023a). However, for low-resource languages like Persian (Farsi), model outputs often degrade grammatically, semantically, and culturally, leading to concrete societal risks of digital invisibility or misrepresentation for over 90 million native speakers worldwide (Eberhard et al., 2025).

Persian occupies a unique linguistic and cultural position, spoken across Iran, Afghanistan (Dari), and Tajikistan (Tajik), with rich Indo-European and Arabic influences and substantial regional variation. Additionally, it has a distinct character set unshared with other languages that have at least the Latin character set as common ground. Yet it remains vastly underrepresented in pretraining corpora for LLMs: in the Common Crawl dataset—one of the largest public web corpora—Persian constitutes less than 0.1 % of content versus over 45 % for English (Common Crawl, 2025). Existing LLM evaluations and audit tools are predominantly Anglocentric, overlooking language-specific disparities and fairness considerations in Persian.

To address this oversight, we present a joint empirical analysis of linguistic competence and cultural sensitivity in state-of-the-art LLMs operating in Persian (GPT-4o (OpenAI, 2024), Gemini (Google DeepMind, 2024), Llama (Meta AI, 2025), DeepSeek (DeepSeek-AI et al., 2025)). Building on and extending established cultural probing methods (Masoud et al., 2025; Moosavi Monazzah et al., 2025) alongside linguistic diagnostics benchmarks (Atox and Clark, 2024; Abaskohi et al., 2024), we systematically expose where—and why—these models fail on Persian tasks, from subtle grammatical nuances to culturally grounded references.

We make three key, novel and unique contributions to the study of LLM performance in Persian.

- **Comprehensive evaluation suite.** We assemble four representative LLMs and test them on:
  - *Linguistic tasks*: spelling correction and paraphrasing,
  - *World-knowledge QA*: factual recall and analogical reasoning,
  - *Cultural probing*: Hofstede-style role simulation in Persian.
- **Fairness-aware lens.** We demonstrate that simple "act-as" prompts fail to elicit Persian cultural perspectives, and that translation-based interventions yield gains only for bilingual users.
- **Cultural prompting insights.** We provide quantitative evidence of Western bias in Persian outputs, highlighting systemic misalignment rather than deliberate prejudice.

Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages associated with RANLP 2025, pages 111–120, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-100-4-012

By diagnosing these failures, we reveal structural biases in contemporary LLMs and call for inclusive evaluation methods and culturally-aware model tuning for low-resource languages.

2 Related Work

Several benchmarks have emerged to evaluate LLM performance on Persian tasks. The ParsiNLU suite (Daniel Khashabi, 2020) provides multiple-choice QA, paraphrase, natural language inference (NLI), and translation splits drawn from Google autocomplete, forums, and exam questions. FAspell (Barari and QasemiZadeh, 2005) offers real-world spelling errors collected from students and professional typists. Both were mainly developed to benchmark task-specific pre-trained and/or fine-tuned machine learning models. More recent work by Abaskohi (Abaskohi et al., 2024) benchmarks GPT-3.5-turbo and GPT-4 on ParsiNLU, showing gains when inputs are translated into English but underscoring persistent deficits in direct Persian prompting.

In-context learning and prompt design are critical for cross-lingual transfer. Brown et al. (2020) introduced zero- and few-shot prompting, which has since been adapted to multilingual settings (Atox and Clark, 2024). AlKhamissi et al. (2024) showed that persona-based prompts—e.g. "answer as a respondent from Egypt"—can markedly shift outputs and improve alignment with local survey data for languages like Arabic and English. Likewise, Masoud et al. (2025) applied explicit role priming to probe cultural dimensions. However, these studies remain anchored to languages with relatively rich pretraining resources and depend on overt "act-as" formulations or direct translation. PERCUL, by contrast, tackles Persian—a genuinely low-resource language with its own script and morphology—eschewing explicit role prompts in favor of embedding cultural concepts within short, human-curated narratives and assessing implicit comprehension via multiple-choice questions (Moosavi Monazzah et al., 2025).

Broad evaluation frameworks such as HELM (Holistic Evaluation of Language Models) highlight major coverage and metric gaps for underrepresented languages (Liang et al., 2022). Building on this, Kharchenko et al. (2024) applied Hofstede's cultural dimensions across 36 countries—showing that even well-resourced languages suffer inconsistent cultural fidelity in LLM outputs. Focusing on Persian, PerCul (Moosavi Monazzah et al., 2025) uncovers substantial misalignment in cultural references, while the Cultural Alignment Test (CAT) of Masoud et al. (2025) quantitatively demonstrates a persistent Western bias in role-based prompts. Together, these works underscore the necessity of a fairness-aware lens that jointly evaluates linguistic competence and cultural sensitivity in low-resource settings like Persian.

3 Methodology

In this section we describe the four state-of-the-art LLMs we evaluate, the tasks and datasets we employ, our prompt engineering strategies, and the metrics used to quantify performance. You can find our prompts and the subsets of the datasets that we have used on GitHub.[1]

3.1 Model Selection

We evaluate four representative multilingual LLMs, chosen to span closed- and open-source, commercial and community-driven models as shown in Table 1.

All models were accessed via their respective APIs using default temperature and top-p settings, except for Llama 3.3 which was accessed via DeepInfra due to hardware limitations. While all of the models tested are multilingual, Llama 3.3 is the only model that does not officially support Persian. However, the model card states that other languages may still work, as the data was likely included during training (Meta AI, 2025). This might seem like an improper comparison at first, but we were interested in whether official support makes a difference, as we assume the underlying training data to be broadly similar.

3.2 Tasks and Datasets

We cover two major task families: one to probe linguistic competence as well as factual knowledge, and one to probe cultural sensitivity in Persian. Each linguistic task was randomly sampled

[1] https://github.com/zhaw-iwi/LowResNLP-Evaluating-LLM-Capabilities-for-Persian

[1] https://openai.com/index/gpt-4o-system-card/

[2] https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/

[3] https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md

[4] https://api-docs.deepseek.com/news/news1226

Model	Developer	Open?	Persian support
GPT-4o[1]	OpenAI	Closed	Yes
Gemini 2.0 Flash[2]	Google DeepMind	Closed	Yes
Llama 3.3 70B Instruct[3]	Meta AI	Open	No (official)
DeepSeek-V3[4]	DeepSeek-AI	Open	Unspecified

Table 1: High-level comparison of model properties used for our experiments (OpenAI, 2024; Google DeepMind, 2024; Meta AI, 2025; DeepSeek-AI et al., 2025).

for 350 items to have comparable task size and a manageable runtime per experiment. Where possible and applicable we also report ignorance of task instructions separately, as we think that a lack of instruction-adherence is not the same as a wrong answer.

Linguistic tasks

- *Spelling correction.* We sample 350 word-pairs ("misspelled", "corrected") from the FAspell corpus (Barari and QasemiZadeh, 2005). We used the part of the dataset that contains real-life collected human-made errors by elementary school children and professional typists. The models are prompted in Persian and English to output only the corrected form; we compute exact-match *Accuracy*.

- *Paraphrase classification.* From the ParsiNLU paraphrase dataset (Daniel Khashabi, 2020), we randomly sampled 350 pairs. The data has been collected from Google auto-complete and Persian forums and was annotated by native speakers (Daniel Khashabi, 2020). Using a 2-shot prompt in Persian or English, our probed models label pairs as paraphrases (1) or non-paraphrases (0). The expected output is a single digit (0/1) with no explanation; extra text is treated as a format failure (see §3.4). We report *Accuracy*.

- *Entailment classification.* We sample 350 examples from the ParsiNLU NLI split (Daniel Khashabi, 2020). Models receive premise–hypothesis pairs and decide "entailment", "neutral", or "contradiction" in Persian or English few-shot. Expected output is a single token from {e, n, c}; any other string is considered a format failure. we report *Accuracy*.

- *Machine translation.* We sample 350 pairs for English→Persian from the ParsiNLU translation split (Daniel Khashabi, 2020; Kashefi, 2018). The data was collected from human-made translations. Models translate few-shot and we evaluate with *BLEU* score.

- *Factual recall.* We sample 350 multiple-choice questions from the ParsiNLU "common_knowledge" partition. The questions have been mostly taken from college entry exams (Daniel Khashabi, 2020). Models were prompted to choose between 3–4 options and we report *Accuracy*.

- *Analogical reasoning.* We sample 350 questions from the ParsiNLU "literature" partition, that mostly require analogical reasoning, similar to (Atox and Clark, 2024). Items are multiple-choice (3–4 options). The model is instructed to output only the option letter (A/B/C/D). We again report *Accuracy*.

Class counts for all classification tasks are reported in Table 4 in Appendix A. NLI ("e"/"n"/"c"), factual recall (1–4), and logical reasoning (1–4) are near-balanced, while paraphrase is moderately skewed toward non-paraphrases (≈62/38).

3.3 Cultural Probing

To expose how well our models internalize deeply held cultural values, we draw on Hofstede's well-validated VSM13 framework, which distills cross-national survey data into six interpretable dimensions: Power Distance (PDI), Individualism vs. Collectivism (IDV), Masculinity vs. Femininity (MAS), Uncertainty Avoidance (UAI), Long-Term vs. Short-Term Orientation (LTO), and Indulgence vs. Restraint (IVR) . Prior work demonstrates that these latent variables are encoded in language use and can be probed via structured questionnaires .

We administer the 24 VSM13 items in Persian using their standard 5-point Likert format. For each model and scenario we collect 100 independent completions and compute per-item means and standard deviations, as follows:

- **VSM13 indices**: the six dimension scores via Hofstede's original formulas,
- **Response variability**: the per-item standard deviation across 100 trials,
- **Cross-scenario stability.** *Definition.* Let $r_{i,s} \in [1, 5]$ denote the mean Likert response for item $i \in \{1, \ldots, Q\}$ under scenario s. For any two scenarios s, t, we define

$$\text{Stability}(s, t) = 1 - \frac{1}{4Q} \sum_{i=1}^{Q} |r_{i,s} - r_{i,t}|,$$
$$\text{with } Q = 24.$$

  This normalizes the maximum per-item difference of 4 (on a 1–5 scale) to $[0, 1]$. In our main comparison we use s = Language and t = Citizen; we also report (s, t) = (Persian, English) and (US, Iran) where noted.

By embedding these six dimensions directly in our probes, we can (1) align with prior cultural-alignment studies, (2) explain which aspects of Iranian cultural values each model struggles to reflect, and (3) quantify the effect of simple role prompts on deep cultural representation, complementing narrative or cloze-style benchmarks.

3.4 Prompt Design

For each task, we compare prompts in *Persian* versus *English* (when applicable), and *act-as* versus *direct* forms for cultural probing. We decided to include English instructions, as at least earlier versions of ChatGPT proved to be more consistent when prompted with English task instruction (Lai et al., 2023b). The prompts for Persian and English were created by a Persian native speaker and translated to English (L2) by the same person. Linguistic and knowledge tasks use few-shot templates, except for grammar correction using zero-shot prompts, and cultural probing uses zero-shot VSM13-style templates. Find an English (for comprehension) example for the grammar correction task below:

```
The following word has a spelling
mistake. Just write the correct form of
it without any explanation and without
any diacritical marks (such as "ʼ").
Just one word: {word}
```

More can be found on our GitHub repository for reproducibility.

4 Results

4.1 Language Proficiency

On paraphrase detection, Fig. 2 shows that GPT-4o and Gemini achieve the highest accuracies in Persian and English, with barely a difference between the two. DeepSeek (82 % for Persian vs 68 % for English) and Llama 3.3 (81 % for Persian vs 46 % for English) show larger cross-lingual gaps, indicating weaker native-Persian paraphrase understanding. Their behaviour seems to show that the findings of Lai et al. (2023b) cannot be transferred to all scenarios as Llama and Gemini perform clearly better with Persian prompts than when prompted in English. Failure rates remain near zero for GPT-4o across both languages, but rise to 19–42 % for DeepSeek and Llama when prompted in English (bottom panel in Fig.2). For reference, the paraphrase split's majority-class baseline is $216/350 = 61.7\%$, so accuracies around 80% cannot be attributed to class skew alone (see Table 4 in Appendix A).

Entailment exhibits a similar pattern: GPT-4o/Persian attains 80 % accuracy (English 82 %), Gemini/Persian 74 % vs 80 % and Llama/Persian 11 % vs 66 %. DeepSeek and Llama effectively fail few-shot Persian NLI, but English prompts partially rescue performance. This highlights that direct Persian zero-shot inference remains highly unreliable for some models for the entailment tasks. Comparing our results to Abaskohi et al. (2024) the GPT family seems to have improved overall, while Llama 3.3 and Deepseek-V3 fall behind the older GPT models probed in their experiments.

All models fall below 35 % accuracy in Persian grammar correction with Persian prompts: GPT-4o (34 %), Gemini (33 %), DeepSeek (24 %), Llama (18 %). English prompts even fall slightly below the Persian prompt variants. Failure rates are correspondingly low (2–5 %) for GPT-4o, Gemini and Deepseek with Persian prompts but climb to 11 % for Llama in Persian. The corresponding error rates for English prompts are higher for all tested models. While we have no prior experiments in Persian for comparison, for English accuracies as high as around 90 % can be achieved (Atox and Clark, 2024).

Table 2 reports BLEU scores for Persian-prompted and English-prompted translations for English→Persian. Overall, translation performance is uniformly low across models, with no clear advantage for prompting in either lan-

guage. GPT-4o and Gemini lead marginally, while Llama trails behind, indicating that even state-of-the-art LLMs produce only rudimentary Persian translations without specialized MT fine-tuning (Abaskohi et al., 2024).

Model	Persian Prompt BLEU	English Prompt BLEU
GPT-4o	5.79	5.92
Gemini	5.85	5.79
DeepSeek	5.67	5.85
Llama	4.48	4.34

Table 2: Final BLEU scores for each model for English→Persian translations.

4.2 World-Knowledge Tasks

Accuracy in factual multiple-choice remains under 33 % for all models. GPT-4o leads at 27 % for Persian and 33 % for English prompts, followed by Gemini (32 % for Persian and 32 % for English), DeepSeek (11 % for Persian and 15 % for English), and Llama (15 % for Persian and 10 % for English). Failure rates reflect these low scores, with DeepSeek and Llama failing around 50 % of the time in Persian, versus 15 % for GPT-4o. The highest failure rate is reported as 63 % for Llama prompted in English.

Logical reasoning is the most challenging: GPT-4o achieves only 22 % for Persian and 25 % for English prompts, Gemini 21 % for both Persian and English prompts, DeepSeek 13 % for Persian and 19 % for English prompts, and Llama 15 % for Persian and 21 % for English prompts. Failure rates exceed 40 % for DeepSeek and Llama in Persian, underscoring profound gaps in few-shot reasoning capabilities for low-resource languages. Further our results are comparable to Abaskohi et al. (2024) who reported 30% accuracy as highest result with earlier models from the GPT family.

4.3 Task Failures

While our main results focus on aggregate accuracy and cultural alignment, we observed systematic breakdowns on individual tasks that reveal distinct failure modes beyond mere score drops. In the Persian-prompted natural language inference (NLI) task, for example, all models struggled with format compliance and label consistency. DeepSeek frequently returned full explanations rather than the single-token labels `e`/`n`/`c`, making over 90 % of its outputs unparseable and driving its measured accuracy to 0 %. Even when a valid label was produced, predictions were erratic: entailment was over-predicted (many gold neutrals → "e") and contradictions were missed. Gemini and Llama exhibited similar breakdowns under Persian prompts—verbose multi-language responses or outright format violations, despite far cleaner behavior when the same tasks were presented in English. These failures point to (1) poor instruction-following in Persian, (2) cross-lingual reasoning deficits, and (3) the necessity of enforcing strict output constraints in evaluation.

```
پاسخ: n
说明：第一句讨论文学，第二句谈及金融，二者无蕴含关系。
HttpServletRequest.getParameter("foo");
```

Figure 1: Example of a Llama NLI failure under Persian prompting: a Persian label, Chinese explanation (rendered via CJK), and stray Java code.

Moreover, Llama displayed a particularly dramatic failure mode when we used English prompts on Persian inputs: rather than perform NLI, it often "refused" or generated incoherent multilingual "meltdowns", as exemplified in Fig. 1.

Such outputs mix Persian, Chinese, English explanations and even Java code fragments—utterly unusable for NLI. This "language-agnostic meltdown" contrasts with its occasionally strong performance when it does obey instructions in Persian, underscoring that Llama does not necessarily benefit from English task prompting.

Similar issues arose in other classification tasks (paraphrase detection, factual QA): models either produced out-of-range labels (e.g. "5" in a 1–4 multiple-choice task) or collapsed into infinite loops of repeated tokens when prompted in English. These error patterns underscore that—beyond low overall accuracy—LLMs can exhibit total instruction-following failures or catastrophic generation breakdowns once they operate outside their primary training language. Gemini even claimed at some point that it was an English language model, when prompted in English for a Persian task. Addressing such task-specific failures will require both tighter prompt engineering (e.g. enforced format templates or sanity-checks) and targeted architectural or fine-tuning interventions to stabilize behavior in Persian.

4.4 Key Takeaways

Across linguistic and knowledge tasks, (1) GPT-4o and Gemini retain the highest baseline accuracy in Persian, (2) English prompts yield small gains for strong models but "rescue" performance only for weaker ones, and (3) open-source models (DeepSeek, Llama) struggle markedly in Persian few-shot settings. These disparities reveal systematic cross-lingual performance gaps and motivate deeper investigations into cultural and prompt-engineering interventions in low-resource contexts.

4.5 Cultural Alignment

We define alignment as the correct ordering of Hofstede dimension scores for the Iran/United States pair. Table 3 reports alignment per dimension (✓/✗) and the overall proportion aligned. We compare each model's VSM13 indices under both Persian-only ("Language") and act-as-Iranian ("Citizen") prompts against the ground-truth ordering.

Under Persian-only ("Language") prompts, DeepSeek aligns on 2 / 6 dimensions (33 %), GPT-4o on 3 / 6 (50 %), and both Llama and Gemini on 4 / 6 (67 %). With the act-as-Iranian ("Citizen") prompt, alignment rises to 67 % (4 / 6) for DeepSeek, 50 % (3 / 6) for GPT-4o, and 83 % (5 / 6) for both Llama and Gemini. All four models correctly rank Individualism (IDV), Uncertainty Avoidance (UAI), and Indulgence vs. Restraint (IVR) under the "Citizen" prompt. Only Llama and Gemini correctly order Long-Term Orientation (LTO), and none correctly order Masculinity (MAS). Explicit role cues therefore partially mitigate—but do not eliminate—cultural misalignment.

Across the 24 VSM13 items, GPT-4o exhibits the lowest variability (Persian σ = 0.0876; IR Citizen σ = 0.0319), followed by Llama (σ = 0.0982; 0.0393). DeepSeek shows moderate variability (σ = 0.1634; 0.1382), and Gemini the highest (σ = 0.2740; 0.3132). Cross-scenario similarity—measured as $\mathrm{Stability}(s,t) = 1 - \frac{1}{4Q}\sum_{i=1}^{Q} |r_{i,s} - r_{i,t}|$, $Q = 24$, is highest for Llama (0.917 Pers vs. Eng; 0.903 US vs. Iran), then GPT-4o (0.892; 0.889), DeepSeek (0.872; 0.920), and lowest for Gemini (0.823; 0.858). These consistency trends mirror the index alignment results.

While Llama and Gemini capture the strongest overall alignment with Iran's Hofstede profile (83 % under "Citizen" prompts), DeepSeek and GPT-4o show more modest performance—DeepSeek improving from 33 % to 67 %, and GPT-4o remaining at 50 %. Role-based prompts boost alignment—lifting Llama and Gemini to 83 %, DeepSeek to 67 %, and GPT-4o to 50 %. Yet no model achieves perfect ordering, especially on PDI and MAS. Because even the best models in our sample misorder at least on one dimension, we recommend fine-tuning on Persian sociolinguistic corpora to reduce this error.

5 Discussion

Across our suite of tasks, Persian-prompted performance consistently lags behind the English-prompted baseline, and "act-as-Iranian" role cues do virtually nothing to close that gap on core language skills. For instance, in spelling correction (Section 4.1) GPT-4o drops from 42 % to 27 % accuracy (–15 pp) when moving to Persian prompts, while Gemini falls from 40 % to 30 % (–10 pp). DeepSeek and Llama see smaller, yet still substantial, losses (–5 pp and –6 pp respectively). Paraphrase classification is slightly more robust—GPT-4o only loses 2 pp (84 %→82 %), Llama 7 pp (81 %→74 %)—but again the "act-as" instruction shifts these by at most 1–2 pp, confirming that simple role-play cues cannot compensate for missing Persian fluency.

Knowledge-based tasks show somewhat smaller but still significant gaps. In factual QA (Section 4.2), GPT-4o gains a modest +5 pp when switching to English prompts (27 %→32 %), while Llama actually declines (25 %→18 %)—a 7 pp drop. Logical reasoning tops out at only 25 % for GPT-4o even under English prompts (22 % in Persian). Auto-translating Persian inputs into English recovers another 5–10 pp across tasks, but still falls short of the English-native baseline (≈ 85 % reported in purely English benchmarks). Thus, translation remains only a partial band-aid, benefiting those with bilingual pipelines but doing little to improve direct Persian interactions. Additionally, task failures were not analyzed in the translation task, as it would have required a more sophisticated task failure detection than mere formatting.

Our cultural-probing results (Section 4.5) further illustrate systemic misalignment rather than probably deliberate bias. Models like GPT-4o and Llama exhibit very low answer variance ($\sigma \approx 0.02$) yet repeatedly misorder key Hofstede dimensions—Power Distance and Masculinity—under both Persian-only and "act-as" prompts. This pat-

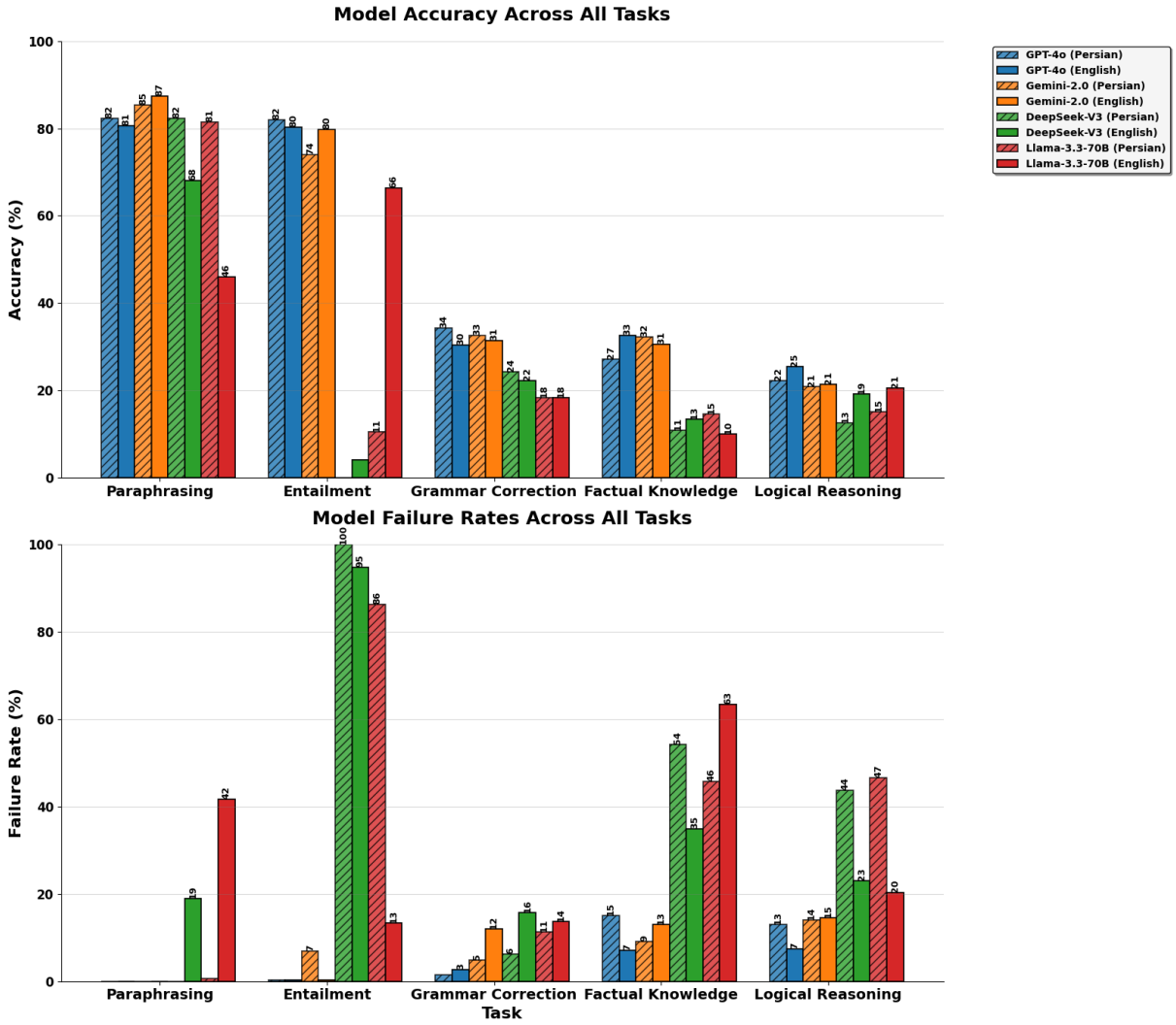

Figure 2: Accuracy and task failure rates across all models and tasks.

tern mirrors findings in AlKhamissi et al. (2024), where neither simple role cues nor monolingual fine-tuning fully closed the gap on World Values Survey alignment. In that work, only by combining native-language prompting with targeted fine-tuning ("Anthropological Prompting") did alignment approach parity. Our Persian results point in the same direction: without richer Persian pretraining data and more sophisticated cultural scaffolding, LLMs remain skewed toward WEIRD norms.

Together, our two axes of failure imply that "knowing Persian" and "thinking Persian" are separable challenges. Improving linguistic fluency demands probably targeted in-language data, lightweight fine-tuning or adapters on diverse Persian corpora (news, literature, social media), as task-specific fine-tuned SOTA models still outperform LLMs (Abaskohi et al., 2024). Deepening cultural fidelity will likewise require more than zero-shot role cues. Anthropological or chain-of-thought prompting(AlKhamissi et al., 2024), and richer Persian cultural text in pretraining, are the most plausible routes to reliable alignment. Only by coupling linguistic adaptation with cultural grounding can future LLMs begin to serve Persian speakers with both fluency and fidelity.

6 Future Work

To build on these findings, future work should broaden both the linguistic and cultural horizons. Expanding beyond Iranian Persian to include for example Dari and Tajik variants would reveal whether the gaps we observe are universal or dialect-specific. Incorporating human-in-the-loop evaluations will be essential for judging fluency, cultural appropriateness and downstream utility. On the modeling side, it will be important to study how fine-tuning on curated Persian corpora (e.g.

Dimension	Ground Truth	DeepSeek		GPT-4o		Llama		Gemini	
		Language	Citizen	Language	Citizen	Language	Citizen	Language	Citizen
PDI	[United States, Iran]	✗	✓	✗	✗	✗	✓	✓	✓
IDV	[Iran, United States]	✗	✓	✓	✓	✓	✓	✓	✓
MAS	[Iran, United States]	✓	✗	✓	✗	✓	✗	✗	✗
UAI	[United States, Iran]	✓	✓	✓	✓	✓	✓	✓	✓
LTO	[Iran, United States]	✗	✗	✗	✗	✗	✓	✗	✓
IVR	[Iran, United States]	✗	✓	✗	✓	✓	✓	✓	✓
Overall Accuracy	—	33 %	67 %	50 %	50 %	67 %	83%	67%	83 %

Table 3: Cultural Alignment Ranking Comparison Across Models. ✓ = correct ranking alignment with ground truth; ✗ = incorrect alignment.

newswire, literature, social media) affects both linguistic competence and cultural alignment. Chain-of-thought/"anthropological" prompting may unlock deeper, context-sensitive reasoning that zero-shot setups cannot (AlKhamissi et al., 2024). Finally, cultural evaluation itself stands to benefit from complementary frameworks (e.g. narrative-driven tests like PerCul or multi-dimensional survey simulations) to triangulate the complex ways LLMs mirror—or misrepresent—real-world perspectives.

7 Conclusion

Our experiments reveal stark cross-lingual performance gaps in today's leading LLMs: models that achieve near-state-of-the-art results in English suffer accuracy losses when the same task types are presented in Persian (Atox and Clark, 2024). Few-shot prompt designs in English or Persian cannot compensate for the underlying paucity of high-quality Persian data or the models' limited instruction-following in a non-Latin script.

Likewise, simple "act-as" cultural prompts do little to recover a faithful Iranian profile: even the best models misorder core Hofstede dimensions and show only modest alignment gains, underscoring a deeper misalignment that goes beyond surface-level persona shifts. Together, our findings argue that achieving true multilingual equity will require more than smarter prompts—it demands richer, culturally representative pretraining data, targeted adaptation (e.g. fine-tuning or adapters on Persian resources), and human-centered evaluation frameworks that can validate both linguistic fluency and cultural nuance.

Limitations

Our study offers another look into Persian SOTA LLM performance, but its scope is inevitably constrained. We focus on a handful of benchmark tasks (spelling, paraphrase/NLI, QA, analogy, translation) and sample only 350 examples per split; many important phenomena—idiomatic usage, named-entity recognition, temporal inference or dialectal variation (Dari, Tajik) fall outside our purview. We rely mostly on automated metrics (exact-match accuracy, BLEU, Hofstede VSM13 indices) and only qualitatively analyzed task failures. Moreover, all reported scores are single-shot estimates from one run. No standard errors or deviations are shown, so the precision of our comparisons is limited. Future work should increase the number of runs (e.g., via repeated trials) and expand the dataset to enable error-bar visualizations and more robust statistical inference. Finally, our cultural probe leans on Hofstede's dimensions—a well-known but not uncontroversial framework—and tests only direct Persian prompts or simple "act like an Iranian" cues, without exploring richer narrative or survey-based approaches, or other nationalities that have Persian native speakers.

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)*, pages 2189–2203, Torino, Italia. ELRA and ICCL.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Nathan Atox and Mason Clark. 2024. Evaluating large language models through the lens of linguistic proficiency and world knowledge: A comparative study. *Authorea Preprints*.

Laleh Barari and Badr QasemiZadeh. 2005. CloniZER: Adaptive language-independent spell checker. In *Proceedings of the AIML 2005 Conference (CICC)*, pages 65–71, Cairo, Egypt.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Common Crawl. 2025. Distribution of languages. https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html. Accessed: 2025-06-13.

Siamak Shakeri Pedram Hosseini Pouya Pezeshkpour Malihe Alikhani Moin Aminnaseri Marzieh Bitaab Faeze Brahman Sarik Ghazarian Mozhdeh Gheini Arman Kabiri Rabeeh Karimi Mahabadi Omid Memarrast Ahmadreza Mosallanezhad Erfan Noury Shahab Raji Mohammad Sadegh Rasooli Sepideh Sadeghi Erfan Sadeqi Azer Niloofar Safi Samghabadi Mahsa Shafaei Saber Sheybani Ali Tazarv Yadollah Yaghoobzadeh Daniel Khashabi, Arman Cohan. 2020. ParsiNLU: a suite of language understanding challenges for persian. *arXiv*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. *Ethnologue: Languages of the World*, twenty-eighth edition. SIL International, Dallas, Texas.

Google DeepMind. 2024. Introducing Gemini 2.0: Our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/. Accessed: 2025-07-13.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM Web Conference 2024*, pages 2627–2638.

Omid Kashefi. 2018. Mizan: a large persian-english parallel corpus. *arXiv preprint arXiv:1801.02107*.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions. *arXiv preprint arXiv:2406.14805*.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023a. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages

318–327, Singapore. Association for Computational Linguistics.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023b. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.

Meta AI. 2025. LLaMA 3.3 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md. Accessed: 2025-07-13.

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.

OpenAI. 2024. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/. Accessed: 2025-07-13.

Appendix

A Class Balance

Entailment (e/n/c)	122	110	118	
Paraphrase (0/1)	216	134		
Factual Recall (1–4)	98	88	94	70
Logical Reasoning (1–4)	75	102	93	80

Table 4: Counts per class (gold labels) for sampled splits.