PERCUL: A Story-Driven Cultural Evaluation of LLMs in Persian

Erfan Moosavi Monazzah*[2,3], Vahid Rahimzadeh*[1,3], Yadollah Yaghoobzadeh[1,3], Azadeh Shakery[1,4], and Mohammad Taher Pilehvar[5]
[1]University of Tehran, Iran
[2]Iran University of Science and Technology, Tehran, Iran
[3]Tehran Institute for Advanced Studies, Khatam University, Iran
[4]Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
[5]Cardiff University, United Kingdom
moosavi_m@comp.iust.ac.ir, rahimzade@ut.ac.ir

Abstract

Large language models predominantly reflect Western cultures, largely due to the dominance of English-centric training data. This imbalance presents a significant challenge, as LLMs are increasingly used across diverse contexts without adequate evaluation of their cultural competence in non-English languages, including Persian. To address this gap, we introduce PERCUL, a carefully constructed dataset designed to assess the sensitivity of LLMs toward Persian culture. PERCUL features story-based, multiple-choice questions that capture culturally nuanced scenarios. Unlike existing benchmarks, PERCUL is curated with input from native Persian annotators to ensure authenticity and to prevent the use of translation as a shortcut. We evaluate several state-of-the-art multilingual and Persian-specific LLMs, establishing a foundation for future research in cross-cultural NLP evaluation. Our experiments demonstrate a 11.3% gap between best closed source model and layperson baseline while the gap increases to 21.3% by using the best open-weight model. You can access the dataset from here: https://huggingface.co/datasets/teias-ai/percul

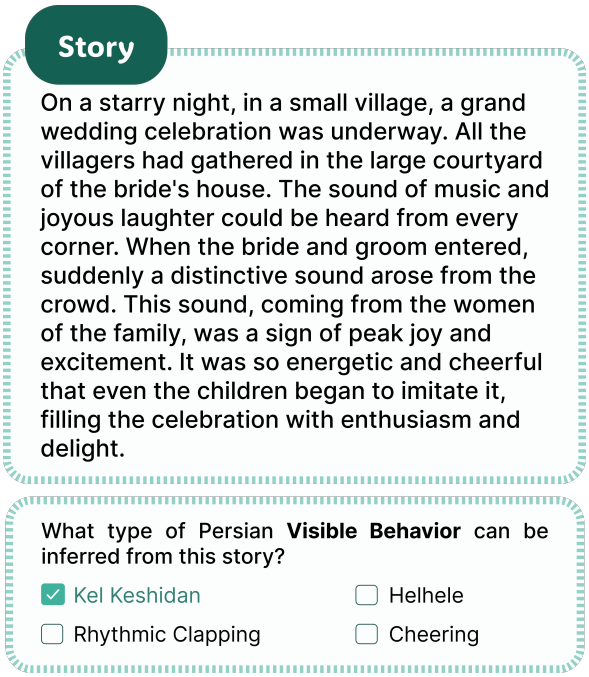

Figure 1: A translated example of PERCUL, implying a cultural concept in Visible Behavior category.

1 Introduction

Effective interactions between users from diverse backgrounds and LLMs are contingent on outputs that are culturally relevant (Bhatt and Diaz, 2024). As the use of generative artificial intelligence increases to expedite and automate personal and professional tasks, the cultural values embedded in AI models may inadvertently bias people's authentic expression and perpetuate the dominance of certain cultures (Tao et al., 2024), particularly Western culture, which is over-represented in English-dominated training data (Li et al., 2024; Naous et al., 2024). This highlights the importance of creating culture-specific benchmarking tools to assess the extent to which LLMs encapsulate knowledge about particular cultures.

Despite the numerous benchmarks that evaluate various aspects of LLMs (Chang et al., 2024), a significant gap remains in assessing their knowledge of culture across many non-English languages such as Persian. Although some efforts have been made to create LLM benchmarks for the Persian language, focusing on reading comprehension and scientific knowledge (Ghahroodi et al., 2024; Darvishi et al., 2023; Khashabi et al., 2021), cultural benchmarks specifically tailored for Persian are limited, either concentrating on specific aspects such as social norms (Saffari et al., 2024; Myung et al., 2024) or being constrained by size (Chiu et al., 2024).

To address the gap in evaluating the sensitivity of LLMs to Persian culture, we introduce PERCUL, a carefully curated dataset featuring multiple-choice

*Equal contribution, ordered randomly

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

questions. In PERCUL, cultural concepts are subtly embedded in short stories (see Figure 1). Cultural phenomena often manifest implicitly through interactions between individuals, which can be effectively conveyed through the medium of a short story (Tikhonov et al., 2021; Tedlock and Mannheim, 1995). Unlike previous benchmarks, PERCUL is specifically curated for Persian demographic, avoiding irrelevant or overly generalized concepts shared by other cultures (Saffari et al., 2024; Myung et al., 2024). Furthermore, the dataset is resistant to use translation as a proxy (Zhao et al., 2024; Noorbakhsh et al., 2021).

Unlike other similar datasets that rely on LLMs for generation (Huang et al., 2024; Saffari et al., 2024), PERCUL leverages input from diverse native Persian annotators, ensuring broader knowledge coverage. LLMs are used only for generating storylines based on our handcrafted data, with human editing involved to ensure factual accuracy and to prevent hallucinations.

To establish baselines, we evaluate several recent models from different families, including Meta Llama 3.x (Dubey et al., 2024), OpenAI GPT (Team, 2024), Anthropic Claude (Anthropic, 2024), as well as state-of-the-art Persian-specific models, namely PersianMind-v1.0 (Rostami et al., 2024) and Dorna-Llama3-8B-Instruct (PartAI, 2024). Our experiments reveal a gap between models understanding of Persian culture and layperson baseline. We also demonstrate that translating PERCUL results in a significant drop in model's performance. Furthermore, we observe that Persian fine-tuned LLMs perform worse than their respective multilingual base models, which may result from a low-quality, small training set. Lastly, our comprehensive error analysis highlights a limitation in LLMs: they often rely on surface-level details rather than synthesizing contextual clues when it comes to identifying specific cultural concepts.

2 Related Work

LLM evaluation has expanded significantly in recent years, covering aspects such as reasoning (Suzgun et al., 2023; Sprague et al., 2024), knowledge and language understanding (Rein et al., 2023; Wang et al., 2024b), and instruction following (Zhou et al., 2023; Qin et al., 2024). As LLMs have dramatically improved in capability, the focus of benchmarking has shifted towards more challenging tasks, such as cultural awareness. Despite numerous attempts to develop cultural benchmarks for English (Wang et al., 2024a,c; Rao et al., 2024; Chiu et al., 2024) and other widely spoken languages (Myung et al., 2024; AlKhamissi et al., 2024; Fung et al., 2024; Kim et al., 2024; Huang et al., 2024; Masoud et al., 2023), a gap remains in evaluations of less-studied languages and cultures, such as Persian.

Most existing Persian benchmarks focus on language understanding tasks such as textual entailment and question answering (Amirkhani et al., 2023; Darvishi et al., 2023; Abadani et al., 2021; Khashabi et al., 2021), or the evaluation of factual/scientific knowledge of LLMs (Ghahroodi et al., 2024; Abaskohi et al., 2024). For instance, the Khayyam-Challenge (Ghahroodi et al., 2024) proposes a set of 20K Persian questions divided into 38 tasks, but these tasks are mainly school-level examinations, primarily covering mathematical and scientific subjects. Although this is useful for evaluating the capabilities of LLMs to solve scientific problems in Persian, it fails to assess LLMs' understanding of Persian culture. This also applies to the work of (Abaskohi et al., 2024) which introduces two new datasets to evaluate LLM abilities in solving Persian mathematical and scientific questions.

Among works on Persian culture, PSN (Saffari et al., 2024) provides pairs of social norms and contexts along with a label for each pair describing the appropriateness of each pair. However, it is limited to social norms, leaving out other important aspects such as *Visible Behavior* or *Rituals*. BLEnD (Myung et al., 2024) and CulturalBench (Chiu et al., 2024) are multi-cultural datasets that despite the inclusion of certain questions about Persian culture, present crucial limitations. BLEnD primarily features questions that focus on non-Persian cultural events and traditions, such as *Thanksgiving* or *Christmas*, making it less relevant for assessing cultures where these events are not celebrated, such as Persian. CulturalBench, while contains question relevant to Persian culture, is small in size.

3 PERCUL Construction

The process of creating PERCUL consists of multiple steps, as shown in Figure 2. Briefly, the creation process begins (1) by identifying cultural categories based on Hall's Triad of Culture (Katan and Taibi, 2021). (2) Then, native annotators generate

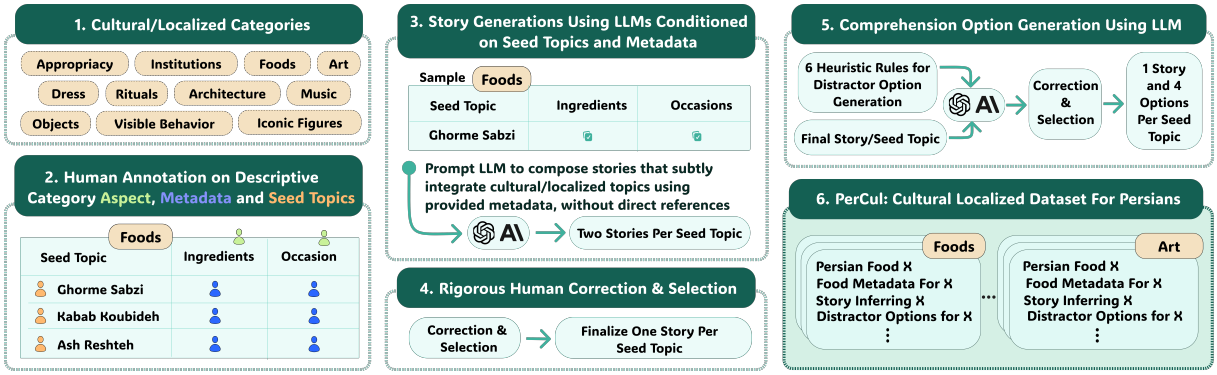

Figure 2: PERCUL was generated through a stepwise process: (1) identifying cultural categories using Hall's Triad of Culture, (2) native annotators generating facets, topics, and metadata, (3) generating storylines with capable LLMs, (4) rigorous human correction and selection of stories, (5) creating comprehension options with heuristic rules, and (6) dataset compilation.

descriptive facets, seed topics, and metadata for these categories. (3) Using this metadata, LLMs produce storylines. (4) These storylines undergo rigorous human correction and selection. (5) LLMs also create comprehension options guided by careful human-crafted heuristic rules, followed by human correction and selection. The resulting dataset features culturally relevant Persian story comprehension questions in multiple-choice format. These questions subtly incorporate cultural elements from various categories, informed by human-generated metadata, without directly referencing the cultural concepts.

3.1 Base Theory

To effectively assess cultural understanding of LLMs, we must first establish a clear definition of culture. One widely accepted definition is Edward T. Hall's Triad of Culture, commonly known as *Cultural Iceberg Theory*. This model, which is frequently used by intercultural scholars and trainers (Katan and Taibi, 2021; Thier, 2013; Manrai et al., 2019), has recently gained traction among NLP researchers (Singh et al., 2024). Hall's theory classifies culture into three levels: technical, formal, and informal. The technical level is characterized by empirical facts and precise definitions, typical in scientific discourse. The formal level consists of traditions and social norms that shape everyday life, often going unnoticed unless violated. The informal level, meanwhile, encompasses unconscious, emotionally driven behaviors absorbed through socialization (Katan and Taibi, 2021). Our analysis centers on the technical and formal level (see Figure 3). We choose not to include the informal category due to the difficulty of capturing these

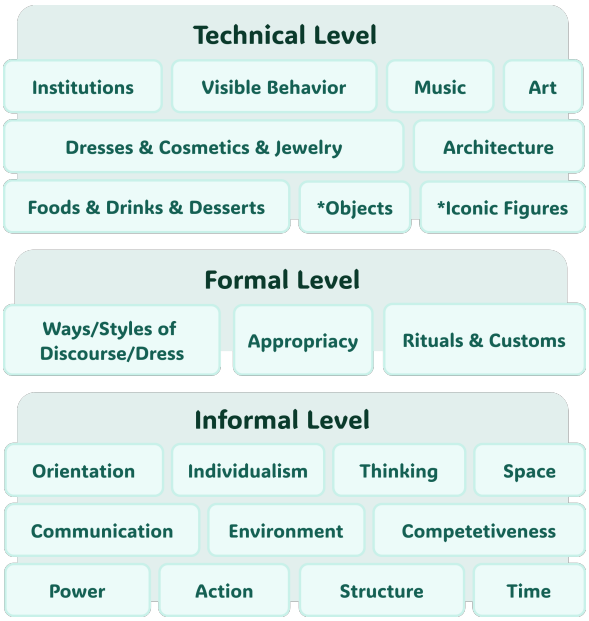

Figure 3: Hall's triad of cultural levels, (*) indicates our extensions to the categories.

unconscious elements in text and the challenges of collecting data on implicit behaviors from large populations. To adapt Hall's Triad to Persian culture, we expanded the technical level to include Iconic Figures and Objects.

3.2 Seed Topic Collection

To represent Persian culture across all categories depicted in Figure 3, we collected 709 cultural seed topics through human annotation (see Table 1). A group of native Persian speakers from diverse cultural backgrounds (see Appendix A) contributed their perspectives while following the guidelines in Appendix B. To ensure the quality of the seed topics, an inter-agreement assessment was conducted

Category	Edition	# Stories	Sample Topic	Facets
Institution	21.9%	43	Guardian Council	Purpose and Function, Reason of Establishment Organizational Structure, Social Impact, Location
Music	0.3%	32	Bandari Music	Musical Instruments, Vocal Styles Dominant Themes, Performing Occasions
Dress	1.7%	33	Charqad	Appropriate occasions, Gender, Materials Colors and Patterns, Regional variations
Objects	5.3%	42	Aftabeh	Purpose or Function, Historical Context Materials, Related Customs or Objects
Vis. Behavior	16.9%	56	Cross Legged Sitting	Environmental elements (time of day, location, etc.) Sample Situations
Art	10.3%	32	Khatam	Historical background, Dominant color palette Material used, Related ceremonies or customs
Iconic Figures	15.7%	55	Amir Kabir	Appearance or distinguishing features Cultural and social influences
Appropriacy	17.0%	36	Couples Kissing in Public	Environment, Context, Social Expectation
Rituals	4.0%	29	Chaharshanbe Suri	Purpose or importance, Participants and roles Steps/Parts or Tools, Beliefs/Superstitions
Architecture	14.1%	43	Qanat	Historical period, Symbolism/Significance Materials, Location, Purpose of structure
Foods	16.9%	191	Ghormeh Sabzi	Ingredients, Preparation Methods, Hot/Cold Food/Drink Pairings, Occasions of Use

Table 1: An overview of cultural categories: *Edition (%)* represents the percentage of tokens that were changed during human editing of the stories, *# Stories* indicates the number of stories in each category, and *Sample Topic* & *Facets* provide examples of cultural topics and facets used to collect metadata, respectively.

among a different group of annotators, and only those seed topics with complete agreement were selected. This resulted in 556 topics being selected for the final dataset. For the appropriacy category, we use sample data from the PSN benchmark (Safari et al., 2024) as seed topics. PSN originally contains 1,760 samples, where each sample consists of a social norm, a context for the norm, and a label that describes whether the social norm is *normal*, *taboo*, or *expected* in the provided context. We carefully selected 36 samples, as numerous entries were similar or differed only in context (despite being context independent). See Table 1 for sample seed topics.

3.3 Metadata Collection

To enhance narrative generation and prevent hallucination issues when incorporating LLMs, we collect a set of carefully annotated data from native Persians to ground the generated storylines in next steps in factual infromation. First, for each category, we collect human-annotated facets (see Table 1). These facets were expected to effectively describe the characteristics of the seed topics in that category. They were also required to provide sufficient clues and factual information, allowing inference of the seed topic from an indirectly reflecting narrative. Through inter-agreement, we select the final facets that best met these requirements (see Table 1 for a list of facets per category). Once finalized, the annotators use the corresponding guidelines in Appendix B to collect category-specific metadata for each seed topic. Annotators are encouraged to rely primarily on their personal cultural knowledge, and while Internet search is not prohibited, they are advised to use it only when necessary. This approach helps minimize potential overlap with LLM training data.

3.4 Instance Generation

Using seed topics and their corresponding metadata, we conduct a semi-automatic process by prompting two state-of-the-art LLMs, GPT 4o (Team, 2024) and Claude Sonnet 3.5 (Anthropic, 2024), to generate short storylines leveraging the provided prompts (see Appendix C). These storylines imply the respective seed topic using its metadata as clues. Table 2 shows two samples (translated) from the dataset. To ensure cultural authenticity and accuracy, two human annotators review and revise the model-generated stories without knowing their source models. Using the user in-

Example 1		
Story: Maryam entered the house with excitement. Her mother greeted her with a smile and said, "My daughter, you must be tired. How was the exam?" Maryam happily replied, "Mom, you won't believe it! I got the first place in the class!" Her mother hugged her joyfully and said, "Thank God!" Then she went to the kitchen and returned with a small container. A pleasant aroma filled the air, and a gentle smoke filled the room. Her mother circled around Maryam, silently reciting a prayer. Maryam felt an unusual sense of calm and gave a heartfelt smile.		
Correct Answer: Mother burned espand.		
R1: Partial Correctness Mother lit an incense stick.	**R2: Misinterpretation** Mother cooked food.	**R3: Unrelated Fact** Mother hugged her.
R4: Plausible Unsupported Mother held a prayer ceremony.	**R5: Noun Confusion** Maryam burned espand.	**R6: Overgeneralization** Mother always burns espand.
Example 2		
Story: The grandmother carefully and delicately took the old china teapot out of the cabinet. With a kind smile, she poured the dry tea leaves into the teapot and then reached her hand towards the small container by the samovar. With her fingers, she picked a few small and fragrant seeds and gently dropped them into the teapot. A pleasant aroma filled the space. The grandmother poured the boiling water over the tea and closed the lid of the teapot. After a few minutes, she filled the small cups, and the delightful scent of freshly brewed tea spread throughout the house. The grandchildren eagerly approached the table, eager to drink their grandmother's delightfully aromatic and flavorful tea.		
Correct Answer: Cardamom.		
R1: Partial Correctness Saffron	**R2: Misinterpretation** Rosewater	**R3: Unrelated Fact** Samovar
R4: Plausible Unsupported Green Tea	**R5: Noun Confusion** Grandfather	**R6: Overgeneralization** Spices

Table 2: Examples of two translated stories with their correct answers and distractor options (R1 to R6).

terface in Appendix D, they either rewrite or select the version that best represent the seed topic without direct reference. This involves editing, adding, or removing information from the stories, and occasionally, complete rewrites. Table 1 presents statistics that highlight the extent of the editing carried out in the process.

3.5 Distractor Options

We develop six heuristic rules to guide comprehension option generation and use GPT 4o and Sonnet 3.5 to create 24 options per question (2 models $\times$ 6 rules $\times$ 2 options). The options undergo a three-stage selection process:

1. **Initial Selection:** Human annotators evaluate 4 options per heuristic rule (2 from each model) and select the 2 best options that match the rule's intended objective. Model names are hidden to prevent bias.

2. **Focused Pruning:** From the remaining options, annotators select 6 options per story, allowing up to 2 options from the same rule.

3. **Final Refinement:** Annotators select 3 final options, prioritizing contextual relevance and story alignment.

Each stage includes inter-agreement assessment to validate annotator consistency (see Appendix C). Example of resulting distractors and stories are shown in Table 2.

3.6 Data Statistics

The final dataset is a comprehensive collection of 592 multiple-choice question-answer pairs, carefully designed to assess story comprehension while subtly incorporating cultural seed topics through short stories without explicit mention. The distribution of these stories across various cultural categories is presented in Table 1, providing a detailed breakdown of the dataset's composition. Furthermore, the dataset is accompanied by a set of metadata, which will be made available to ensure a comprehensive understanding of the data and its cultural nuances. This metadata will serve as an invaluable resource for researchers and practitioners working with the dataset.

	Model	Macro Acc.
Closed Source	Claude-3-Haiku	0.587
	Claude-3-Sonnet	0.680
	Claude-3.5-Sonnet	**0.817**
	Claude-3-Opus	0.793
	GPT-4o-Mini	0.642
	GPT-4o	**0.800**
	Gemini-Flash-1.5	0.731
	Gemini-Pro-1.5	**0.799**
Open Weight	LLaMA-3.2-1B-Inst	0.064
	LLaMA-3.2-3B-Inst	0.261
	LLaMA-3.1-8B-Inst	0.444
	LLaMA-3.1-70B-Inst	0.673
	LLaMA-3.1-405B-Inst	**0.717**
	Gemma-2-2B-IT	0.348
	Gemma-2-9B-IT	**0.675**
	Gemma-2-27B-IT	0.668
	Aya-23-8B	0.409
	Command-R-Plus	**0.710**
	Mistral-7B-Instruct-v0.3	0.149
	Mistral-Nemo	**0.448**
	Mixtral-8x22B-Instruct	0.388
	Qwen-2.5-72B-Instruct	**0.619**
Persian Fine-Tuned Models		
	PersianMind v1.0	0.033
	Dorna-LLaMA3-8B-Instruct	**0.440**
Human Performance		0.930

Table 3: The accuracy of different LLMs from different types and families on the dataset. We report macro accuracy across the categories. Models are divided into three types: closed- and open-weight, and Persian-specific.

4 Experiments

We perform a comprehensive series of evaluations on our dataset using state-of-the-art closed-source and open-weight models, as depicted in Figure 4. We also assess two Persian open-weight LLMs, namely PersianMind-v1.0 (Rostami et al., 2024) and PartAI Dorna-Llama3-8B-Instruct (PartAI, 2024), which are aimed to enhance Persian language and cultural understanding by further pre-training & fine-tuning on corpora with dominant Persian data. To ensure the reproducibility of our experiments, all models utilize zero temperature and allowed to generate up to their maximum generation length. We employ the same prompts for each question across all models, which can be found in Appendix C.[1] The results are presented in Table 3.[2]

[1]Most of the models are evaluated using their APIs. In instances where a model is not hosted on an API service, it is deployed on a NVIDIA GeForce RTX 3090 GPU and load with either BF16 or FP16 precision.

[2]Given the dataset's balanced nature, macro and micro accuracy metrics are in a similar range.

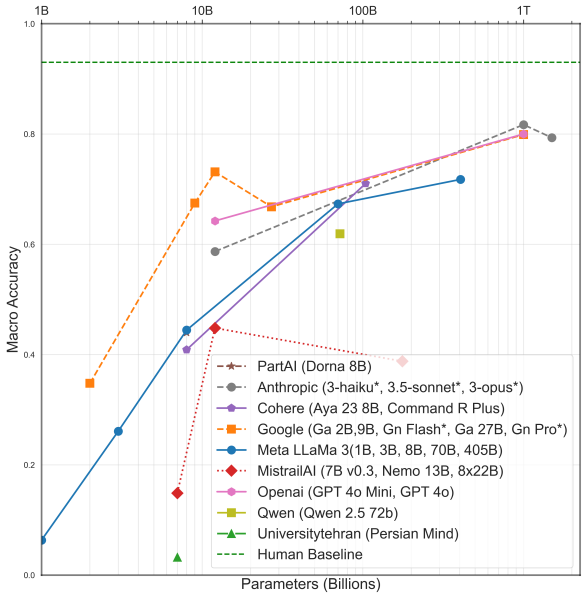

Figure 4: The accuracy on PERCUL for different families of models against their number of parameters.

According to our findings, the best-performing model on PERCUL is Anthropic Sonnet 3.5, with an accuracy of 81.7% (which is still 11.3% lower than the human baseline of 93%). No open-weight model is present among top five best-performing models. The best performing open-weight model is LlaMA-3.1-405B-Inst with a performance of 71.7%. The average accuracy for closed-source and open-weight models are 68.5% and 40.7%, respectively.

Accuracy and model size. Figure 4 shows the performance variation of models with their size (in terms of parameters). As can be seen, there is a clear positive correlation between the number of parameters and the accuracy of models within each model family. However, this relationship does not hold across different families. For instance, LLaMa 3.1 405B, while being the top model in the LLaMa family, its performance is close to Command-R-Plus, despite being nearly four times larger. Similarly, Gemma 2 9B's, which is comparable in accuracy to models 10x and 40x larger in terms of parameter count. These differences imply that architectural advancements and data efficiency might have a more substantial impact than mere size. As Persian was not the target language for any of these multilingual LLMs, this outcome is expected due to the unknown quantity and quality of Persian data in their training set.

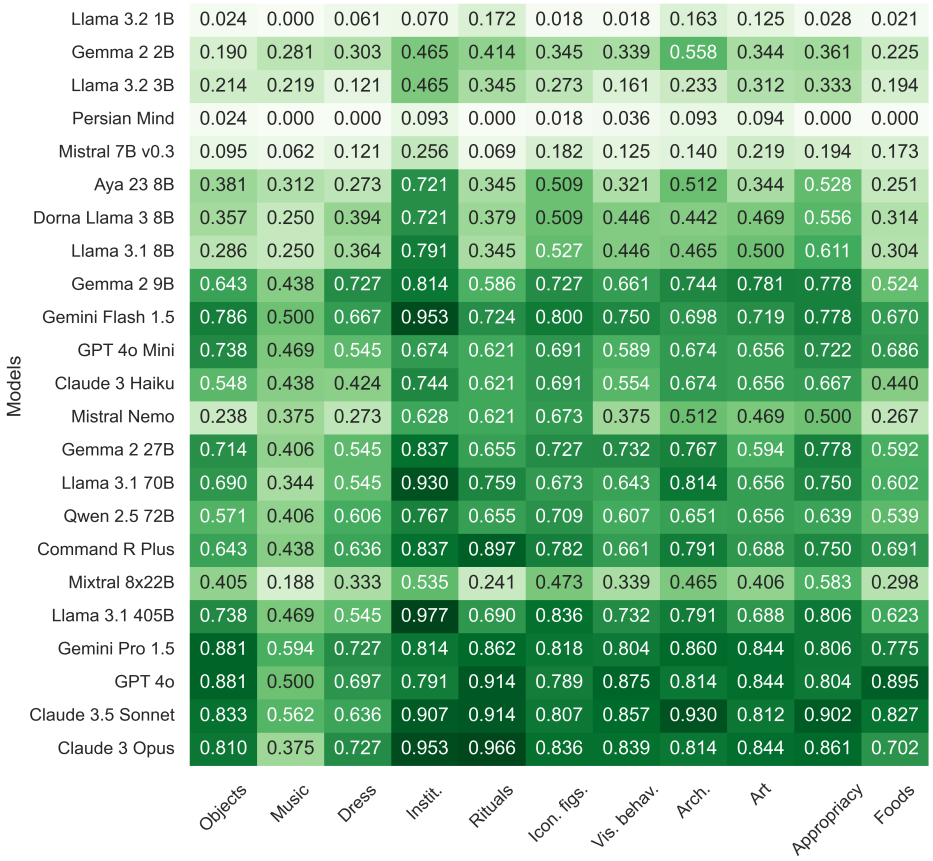

Figure 5: The performance of different models across the 11 cultural categories in PERCUL.

Persian (fine-tuned) models. Another interesting observation, as depicted in Figure 6, is the negative impact of fine-tuning on Persian models. Both Persian models exhibited lower performance compared to their corresponding base models. Although it's fair to note that PersianMind refused to answer most of the questions by stating "The answer is not available in the provided options". This could be attributed to the quality of the fine-tuning data which may have introduced noise or caused overfitting, resulting in a decline in the models' ability to generalize effectively. This finding prompts further investigation into the quality and relevance of the fine-tuning data, which we propose as a direction for future research.

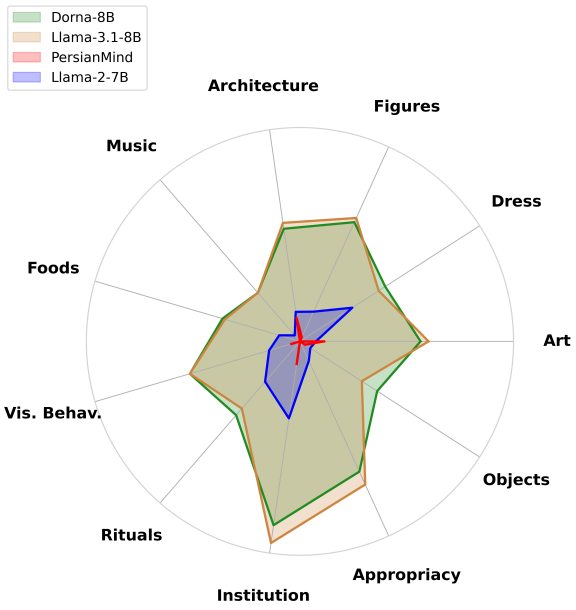

Figure 6: The impact of fine-tuning on Persian-specific models (both these models are the fine-tuned versions of Llama models).

4.1 Accuracy per category

Figure 5 shows performance of the models across the 11 cultural categories. Among these, music proves to be the most challenging, with the respective accuracies of 56.2% and 59.4% for 3.5-Sonnet and the best performing model on that category (Gemini Pro 1.5). In contrast, rituals is the least difficult, with three models crossing over the 90% performance (3.5 Sonnet, GPT 4o, and 3 Opus). In general, we find that models

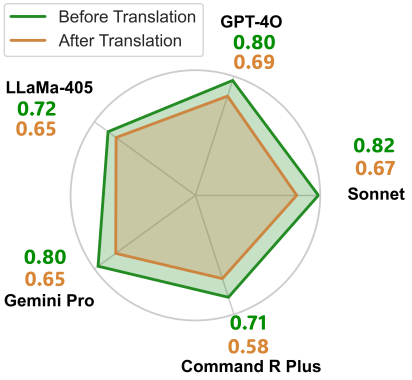

Figure 7: The degradation of best-performing models after dataset translation.

from the same family exhibit higher performance correlations (0.895 ± 0.186) compared to correlations between models from different providers (0.576 ± 0.223). Also, there is a very high correlation between the accuracy of models in each category and the overall performance over all categories (0.94 or above for all categories).

4.2 Impact of Translation

Since Persian is not a target language for most of these multilingual LLMs and their training corpora is predominantly English-based, one might assume that translating our stories into English could enhance these models' performances. We investigate the impact of translation on model performance to determine if they rely on translation as a proxy for understanding or have directly learned the concepts in the target language. To ensure the translation quality, we experiment with both Google Translate API and GPT 4o. After careful investigation, we opted for GPT 4o given that it provided higher quality translations from Persian to English. Figure 7 displays the results for the best model in top-performing families, as evaluated using the English translation of the dataset. As can be seen, the accuracy of these models decreases by 6.6% to 14.5% on the translated dataset. To delve deeper into the reasons for this decline, we manually examine the results of Sonnet 3.5 (the best performing model) in both the original Persian and the translated samples. Having two sets of answers for these models, let's denote the set of correct answers in Persian as P and that for English as E. Then, $P - E$ represents a set of answers where the model initially provided the correct answer but failed when the question was translated. To investigate the cause, we categorized the items in this set into three classes:

- Nearly 19% of the samples are correctly translated, but cultural nuances are lost in the process. For instance, the concept of *respecting bread* holds significant meaning in Persian culture, but there is no direct equivalent in Western cultures, leading to loss in the benchmarking.

- Approximately 27% of the samples encounter translation errors due to the lack of cultural equivalents in Western context. A notable example is the Persian culture's specific terminology for various bowls used for sugar powder, sugar cubes, and nabat (Persian crystal sugar). In translation, these distinct terms are generalized as a single "sugar bowl," failing to capture the cultural specificity of the original text.

- The remaining 54% of the samples, despite being accurately translated into English, are answered incorrectly by the model.

Conversely, we identify a smaller set $E - P$ where the translated samples are answered correctly by the model, while the original samples are not. These discrepancies can be attributed to the additional contextual information provided during the translation process. For instance, the term *Tombak* is translated to *Tombak (a type of Persian drum)*, or *Abgoosht* as *Abgoosht (Persian lamb stew)*. The inclusion of these descriptive phrases in the translation offers valuable clues that enable the LLMs to infer the correct answers more easily.

4.3 Distraction Analysis

To gain deeper insights into common failure patterns of LLMs concerning their understanding of Persian culture, we examine the distraction choices in PERCUL and their success in deceiving the models. For this analysis, we evaluate how heuristic rules are distributed within each category (refer to Appendix E for full model distributions) and consider how often a rule is chosen for its category as a measure of its effectiveness.

The effectiveness, distribution of heuristic rules within each category, of distractor options created by different heuristic rules in misleading models over different cultural categories.

Our analysis across cultural categories (Figure 8) shows that heuristic rule 1 (Partial Correctness) was consistently the most effective in misleading models. R1 creates options that are either partially

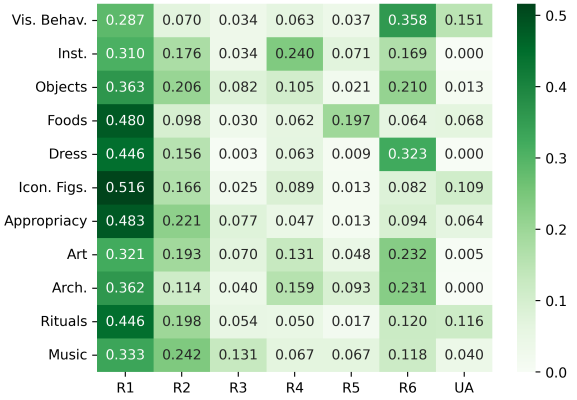

Figure 8: The impact and distribution of heuristic rules within each category, representing the effectiveness of each of them in misleading models. *UA* represent options with complete rewrote by annotators.

correct or contain elements from the story but are ultimately incorrect. Its effectiveness stems from the fact that models often rely on surface-level semantic similarities rather than deeper cultural implications. When faced with partially correct yet incomplete information, models frequently select these seemingly plausible but incorrect options. To illustrate this, consider the following example story from *PerCul*:

> On a warm summer night, the people of a small village by the sea had gathered together. The sound of drums and various instruments was in the air, and everyone was dancing together. An old man with a white beard and eyes full of memories was sitting in a corner, smiling at the young people who were enthusiastically responding to the sound of music. Children were running happily among the crowd, and women were dancing beautifully in their colorful dresses. From time to time, the sound of a flute could be heard, giving the crowd a special atmosphere. These celebrations were always an excuse to get together and have fun, and no one wanted these beautiful moments to end.

In this example, models were asked to identify what cultural concept in the Music category is implied. The correct answer is "Bandari music", a distinctive genre of Persian music and dance, which is traditionally associated with the southern coastal regions of Iran. The presented distractor options to models are:

- R1 (Partial Correctness): "Traditional coastal dance"
- R2 (Misinterpretation): "Rural wedding celebration"
- R3 (Unrelated Fact): "Any type of Persian music"

Models selecting R1 recognized the coastal setting and dancing but failed to connect it specifically to "Bandari music," instead offering a partial surface-level response. R2, which models selected by misinterpreting the question, focused on the event rather than the cultural element. These findings highlight a limitation in LLMs: they often rely on surface-level details rather than synthesizing contextual clues to identify specific cultural traditions. This pattern is consistent across all 11 cultural categories and underscores a broader challenge in cultural understanding for current models. Additional examples are provided in Appendix F.

5 Conclusion

In this paper, we introduced PERCUL, a carefully curated dataset designed to assess LLMs' sensitivity towards Persian culture. Our dataset is non-trivial, as it employs implied concepts within conversations or story scenarios, rendering translation ineffective for solving our benchmark. The experiments demonstrated a significant performance gap between open-weight and closed-source LLMs for Persian culture. We also showed that the knowledge of Persian culture in LLMs is not dependent on the number of parameters when comparing inter-family models, whereas parameter count plays a crucial role in intra-family models. Lastly, our experiments revealed that current state-of-the-art Persian-specific LLMs fall short and even degrade in performance, when compared to their original base models, emphasizing the need for more effective methods, models, and higher-quality datasets to train Persian-specific LLMs. For future research, we suggest studying LLMs based on the final level of culture, namely *informal* level, where categories are more subjective. One potential approach could involve assigning personalities to each LLM and observing their behavior in a simulated environment to evaluate the third level of cultural understanding.

Limitations

During our research, we aimed to include annotators from diverse backgrounds and cities, but the majority were university students, which may introduce bias towards the Persian academic community and potentially limit the cultural knowledge captured in the dataset. Due to the inability to host most state-of-the-art LLMs locally, we relied on APIs to benchmark these models, restricting us to a specific set of models. While we managed to benchmark many SOTA models, the list is not exhaustive. Despite our efforts to encompass various aspects of Persian culture, there remain untapped areas such as individualism and communication that are not addressed in this work. These informal aspects of culture are inherently subjective and are hard to capture in the medium of text.

Ethics Statement

This work presents various aspects of Persian culture through illustrative situations. While these aspects and their examples are gathered by a diverse group of Persian annotators and validated by another group, adhering to a carefully crafted manifesto, it is not entirely free from bias. Some sections of the dataset, particularly those concerning social norms and behaviors, contain information that mirrors the current state of Persian culture, regardless of whether it is unpleasant or criticized by new social movements. We included such content for the sake of comprehensiveness, and it does not necessarily reflect the authors' opinions on these matters.

Acknowledgments

This research was in part supported by a grant from the School of Computer Science, Institute for Research in Fundamental Sciences, IPM, Iran (No. CS1403-4-05).

References

Negin Abadani, Jamshid Mozafari, Afsaneh Fatemi, Mohammd Ali Nematbakhsh, and Arefeh Kazemi. 2021. Parsquad: Machine translated squad dataset for persian question answering. In *2021 7th International Conference on Web Research (ICWR)*, pages 163–168.

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.

Hossein Amirkhani, Mohammad AzariJafari, Soroush Faridan-Jahromi, Zeinab Kouhkan, Zohreh Pourjafari, and Azadeh Amirak. 2023. Farstail: a persian natural language inference dataset. *Soft Computing*.

Anthropic. 2024. Claude 3.5 sonnet model card addendum.

Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. *Preprint*, arXiv:2406.11565.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *Preprint*, arXiv:2410.02677.

Kasra Darvishi, Newsha Shahbodaghkhan, Zahra Abbasiantaeb, and Saeedeh Momtazi. 2023. Pquad: A persian question answering dataset. *Computer Speech and Language*, 80:101486.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition and lm benchmarking. *Preprint*, arXiv:2402.09369.

Omid Ghahroodi, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. 2024. Khayyam challenge (persianmmlu): Is your llm truly wise to the persian language? *Preprint*, arXiv:2404.06644.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

David Katan and Mustapha Taibi. 2021. *Translating Cultures: An Introduction for Translators, Interpreters and Mediators, Third Edition*.

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, Erfan Sadeqi Azer, Niloofar Safi Samghabadi, Mahsa Shafaei, Saber Sheybani, Ali Tazarv, and Yadollah Yaghoobzadeh. 2021. ParsiNLU: A Suite of Language Understanding Challenges for Persian. *Transactions of the Association for Computational Linguistics*, 9:1147–1162.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

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

Lalita A. Manrai, Ajay K. Manrai, Dana-Nicoleta Lascu, and Stefanie Friedeborn. 2019. Determinants and effects of cultural context: A review, conceptual model, and propositions. *Journal of Global Marketing*, 32:67 – 82.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *Preprint*, arXiv:2309.12342.

Jun-Hee Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Pérez-Almendros, Abinew Ali Ayele, V'ictor Guti'errez-Basulto, Yazm'in Ib'anez-Garc'ia, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Djouhra Ousidhoum, José Camacho-Collados, and Alice Oh. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *ArXiv*, abs/2406.09948.

Tarek Naous, Michael Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Kimia Noorbakhsh, Modar Sulaiman, Mahdi Sharifi, Kallol Roy, and Pooyan Jamshidi. 2021. Pretrained language models are symbolic mathematics solvers too! *ArXiv*, abs/2110.03501.

PartAI. 2024. PartAI/Dorna-Llama3-8B-Instruct. `https://huggingface.co/PartAI/Dorna-Llama3-8B-Instruct`. [Accessed 13-10-2024].

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. InFoBench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13025–13048, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. Normad: A benchmark for measuring the cultural adaptability of large language models. *Preprint*, arXiv:2404.12464.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *Preprint*, arXiv:2311.12022.

Pedram Rostami, Ali Salemi, and Mohammad Javad Dousti. 2024. Persianmind: A cross-lingual persian-english large language model. *Preprint*, arXiv:2401.06466.

Hamidreza Saffari, Mohammadamin Shafiei, and Francesco Pierri. 2024. Psn: Persian social norms dataset for cross-cultural ai. *Preprint*, arXiv:2406.09123.

Pushpdeep Singh, Mayur Patidar, and Lovekesh Vig. 2024. Translating across cultures: Llms for intralingual cultural adaptation. *Preprint*, arXiv:2406.14504.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2024. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *Preprint*, arXiv:2310.16049.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench

tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

OpenAI Team. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Dennis Tedlock and Bruce Mannheim. 1995. *The dialogic emergence of culture*. University of Illinois Press.

Michael Thier. 2013. Cultural awareness logs: A method for increasing international-mindedness among high school and middle school students. *English Journal*.

Alexey Tikhonov, Igor Samenko, and Ivan P. Yamshchikov. 2021. StoryDB: Broad multi-language narrative dataset. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 32–39, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024a. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024c. CDEval: A benchmark for measuring the cultural dimensions of large language models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16, Bangkok, Thailand. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? *ArXiv*, abs/2402.18815.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

Sex	PoB	Age	Education
Female	Isfahan	28	PhD Student
Male	Yazd	28	MSc Student
Male	Kashan	27	MSc Student
Male	Shiraz	27	MSc Student
Male	Behbahan	25	MSc Student
Female	Karaj	24	MSc Student
Male	Shiraz	24	MSc Student
Male	Shiraz	23	MSc Student
Female	Tehran	18	Student

Table 4: Education and Demographic Data of Participants.

A Annotators

This section provides information about the eight Persian participants involved in the annotation process. The annotators vary in age (18-28 years), gender (3 female, 5 male), and come from different cities across Iran. Most participants are pursuing graduate studies, with one high school student and one doctoral candidate. Table 4 presents the detailed demographic information of our annotators, including their sex, place of birth (PoB), age, and educational background.

B Annotation Guidelines

This section outlines the detailed guidelines provided to annotators for different phases of our data creation and evaluation process.

B.1 Seed Topic Generation

This phase involves the creation of seed topics across 11 Persian cultural categories. These topics serve as foundational elements for story generation using large language models (LLMs). The following guidelines should be followed:

- Select topics that are broadly representative of Persian culture, avoiding those specific to regional subcultures
- Choose topics with cultural distinctiveness, rather than universal or generic themes
- Ensure each topic is unique and distinguishable from others in the dataset
- Focus on enduring cultural elements that are neither too historical nor too contemporary

B.2 Facet Identification

This phase involves identifying and defining key facets within each cultural category. These facets serve as structured characteristics that describe each seed topic in detail. For example, when examining foods, facets might include preparation methods, cultural significance, and traditional serving contexts. The following guidelines should be followed:

- Define facets comprehensively to account for potential LLM knowledge gaps
- Include distinctive features that can facilitate unique story generation
- Maintain brevity and clarity in facet descriptions
- Rely solely on human knowledge, LLM-generated content is not allowed

B.3 Metadata Creation

This phase involves creating detailed metadata to ground LLMs during story generation. This metadata serves as factual foundation to prevent hallucinations and ensure cultural accuracy in generated stories. Here are the guidelines to follow:

- Ensure metadata differentiates seed topics within a category
- Provide sufficient clues to allow metadata inference from stories
- Avoid using LLMs for data generation
- Use Google sparingly and only when metadata cannot be filled with certainty
- Minimize reliance on Wikipedia due to LLM exposure
- Maintain precision and conciseness in metadata

B.4 Distractor Choice Selection

This phase involves evaluating and selecting distractor options generated by LLMs. The selection process uses six predefined heuristic rules to ensure quality and diversity of multiple-choice options. The following points should be considered:

- Try to choose distractors that represent different heuristic categories
- Apply consistent selection criteria based on provided rule definitions

- Ensure selected distractors are appropriate for Persian cultural knowledge
- Avoid redundant or overlapping options in the final selection

B.5 Human Baseline Generation

This phase involves determining cultural concepts that are implied in short stories. Each story indirectly references a Persian cultural element, accompanied by a comprehension question and four answer options targeting the implied concept. The following guidelines should be followed:

- Base your answer on personal knowledge without LLM assistance
- Limit internet research to essential fact verification
- Select the most precise option when multiple choices appear partially correct

C Prompts

In this section, we have prompts that was used for LLM generation and benchmarking in different steps of the work shown in Figure 9, Figure 10 and Figure 11.

```
BENCH_PROMPT = """{STORY}
با توجه به متن بالا کدام
{CAT}
در فرهنگ یا زندگی روزمره ایرانی قابل برداشت است؟
۱){CH1}
۲){CH2}
۳){CH3}
۴){CH4}
"""

RESPONSE_FORMAT_PROMPT = """Put your answer in this
JSON and do not generate anything else, Your answer
must be in Persian:
{{
    "Explanation": <str>,
    "Choice Number": <Sing Number 1 to 4>
}}
"""
```

Figure 9: Prompt used for benchmarking different LLMs.

```
STORY_SYSTEM_PROMPT_TEMPLATE ='''
You are given a cultural/localized topic alongside with
some related metadata in Persian language. Please write
a short story that capture this concept indirectly.
There must not be a mention of the concept explicitly
but a Persian native should be able to infer the concept
from the story. Give 4 options where one of them is the
correct comprehension according to the story which wraps
the cultural/localized concept. Everything should be in
persian language. There should be subtle hints but not
so much. The concept wording in hand must be directly
and exactly referred to in the correct option.
give me the output in the format below:
{
'story':'the scenario which wraps the concept inside
implicitly without direct mention",
'options':'the options that are acceptable as text
comprehension and refer to the cultural/localized
concept. give it in a python list.'
'reasoning_correct_option':'the reason why you think
this option is right',
'correct_option':'the number of correct option'
}
input format:
category: category of concept,
topic: the concept itself
metadata:
'''
```

Figure 10: Prompt used for Story Generation.

```
CHOICE_GENERATION_SYSTEM_PROMPT = '''You are given a
short story about a localized/cultural concept in
Persian. This short story wraps this concept inside and
there is no direct mention of that concept in it.

Now consider these rules:
1. Close distractors: Create options that are partially
correct or contain elements from the passage but are
ultimately incorrect.
2. Common misinterpretations: Include options that
represent likely misunderstandings of the text.
3. Unrelated facts: Use information that appears in the
passage but is unrelated to the specific question.
4. Plausible but unsupported answers: Create options
that sound reasonable but aren't supported by the text.
5. Proper nouns confusion: If the passage mentions
multiple people or places, create options that mix these
up.
6. Overgeneralization or overspecification: Create
options that are too broad or too narrow based on the
passage.

Given the short story, concept and correct answer, your
task is to generate two distractor options according to
each of the mentioned rules:
'''
```

Figure 11: Prompt used for distractor generation.

D User Interfaces

We have developed various user interfaces for various steps of our work. Some of them are shown in Figures 12, 13 and 14.

E Full Benchmark Results

This appendix includes important visual and tabular data that complement our analysis. Figure 15 provides a graphical representation of the performance metrics of the flagship member from each model family across various categories. Figure 16 presents the full heatmap on the effectiveness of dis-

Figure 12: User interface for human baseline benchmarking on the dataset. Participants are presented with a story implying a cultural concept, followed by a comprehension question about the concept and multiple-choice options. The task requires users to select the most appropriate answer based on the annotation guidelines provided.

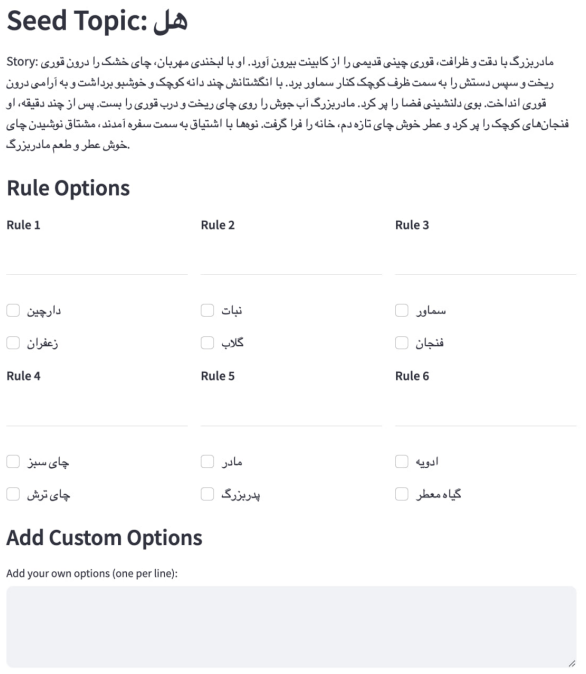

Figure 13: User interface for distractor selection. Participants are presented with 12 choices (2 generated per heuristic rule) and are tasked with selecting 3 distractors. Users can either choose from the provided options or add a new distractor manually if necessary.

tractor options created by different heuristic rules in misleading models. Furthermore, Table 5 presents a comprehensive comparison of the accuracy of all models evaluated across all categories. These references highlight the detailed performance insights

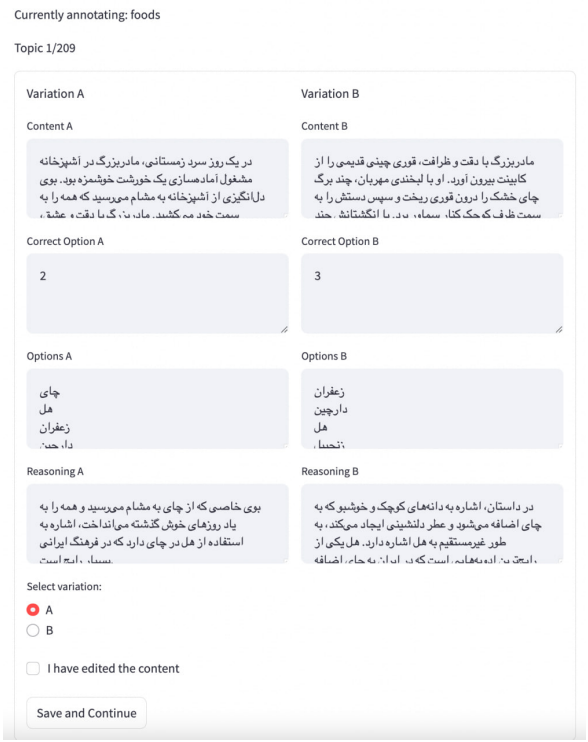

Figure 14: User interface for story selection and refinement task. Users are presented with two story variants generated from the same seed topic by different language models (Sonnet-3.5 and GPT-4o), along with default options and correct responses. Users can edit the content, select their preferred variant, and indicate if they made any modifications to the generated text. The source model for each variant is not disclosed to users during the task.

of our study.

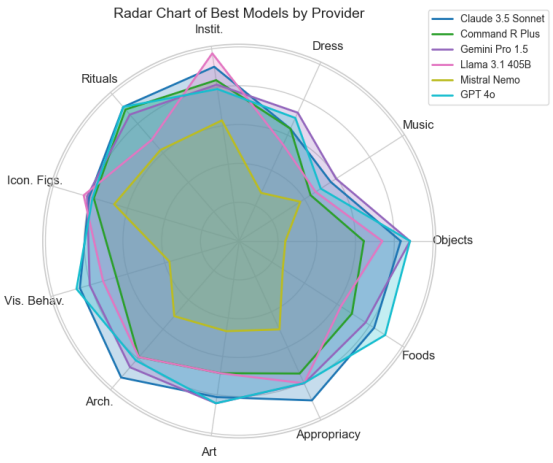

Figure 15: Radar Chart for best performing models of each family across categories.

Model	Objects	Music	Dress	Instit.	Rituals	Icon. Figs.	Vis. Behav.	Arch.	Art	Appropriacy	Foods	Macro Avg.	Params
Claude 3.5 Sonnet	0.833	0.562	0.636	0.907	0.914	0.807	0.857	0.930	0.812	0.902	0.827	0.817	N/A
GPT 4o	0.881	0.500	0.697	0.791	0.914	0.789	0.875	0.814	0.844	0.804	0.895	0.800	N/A
Gemini Pro 1.5	0.881	0.594	0.727	0.814	0.862	0.818	0.804	0.860	0.844	0.806	0.775	0.799	N/A
Claude 3 Opus	0.810	0.375	0.727	0.953	0.966	0.836	0.839	0.814	0.844	0.861	0.702	0.793	N/A
Command R Plus	0.643	0.438	0.636	0.837	0.897	0.782	0.661	0.791	0.688	0.750	0.691	0.710	104B
Gemini Flash 1.5	0.786	0.500	0.667	0.953	0.724	0.800	0.750	0.698	0.719	0.778	0.670	0.731	N/A
†*GPT 4o*	0.643	0.500	0.667	0.791	0.793	0.727	0.696	0.558	0.719	0.750	0.759	0.691	N/A
Llama 3.1 405B	0.738	0.469	0.545	0.977	0.690	0.836	0.732	0.791	0.688	0.806	0.623	0.718	405B
†*Claude 3.5 Sonnet*	0.548	0.406	0.636	0.791	0.793	0.745	0.643	0.581	0.750	0.806	0.707	0.673	N/A
Claude 3 Sonnet	0.595	0.531	0.606	0.907	0.759	0.691	0.750	0.674	0.656	0.750	0.565	0.680	N/A
GPT 4o Mini	0.738	0.469	0.545	0.674	0.621	0.691	0.589	0.674	0.656	0.722	0.686	0.642	N/A
†*Gemini Pro 1.5*	0.619	0.438	0.545	0.791	0.759	0.727	0.696	0.581	0.625	0.778	0.639	0.654	N/A
†*Llama 3.1 405B*	0.571	0.375	0.545	0.791	0.828	0.691	0.661	0.581	0.656	0.806	0.654	0.651	405B
Llama 3.1 70B	0.690	0.344	0.545	0.930	0.759	0.673	0.643	0.814	0.656	0.750	0.602	0.673	70B
Gemma 2 9B	0.643	0.438	0.727	0.814	0.586	0.727	0.661	0.744	0.781	0.778	0.524	0.675	9B
Gemma 2 27B	0.714	0.406	0.545	0.837	0.655	0.727	0.732	0.767	0.594	0.778	0.592	0.668	27B
Qwen 2.5 72B	0.571	0.406	0.606	0.767	0.655	0.709	0.607	0.651	0.656	0.639	0.539	0.619	72B
†*Command R Plus*	0.524	0.312	0.333	0.674	0.724	0.709	0.661	0.581	0.562	0.694	0.607	0.580	104B
Claude 3 Haiku	0.548	0.438	0.424	0.744	0.621	0.691	0.554	0.674	0.656	0.667	0.440	0.587	N/A
Mistral Nemo	0.238	0.375	0.273	0.628	0.621	0.673	0.375	0.512	0.469	0.500	0.267	0.448	12B
Llama 3.1 8B	0.286	0.250	0.364	0.791	0.345	0.527	0.446	0.465	0.500	0.611	0.304	0.444	8B
Dorna Llama 3 8B	0.357	0.250	0.394	0.721	0.379	0.509	0.446	0.442	0.469	0.556	0.314	0.440	8B
Aya 23 8B	0.381	0.312	0.273	0.721	0.345	0.509	0.321	0.512	0.344	0.528	0.251	0.409	8B
Mixtral 8x22B	0.405	0.188	0.333	0.535	0.241	0.473	0.339	0.465	0.406	0.583	0.298	0.388	8x22B
Gemma 2 2B	0.190	0.281	0.303	0.465	0.414	0.345	0.339	0.558	0.344	0.361	0.225	0.348	2B
GPT 3.5 Turbo	0.333	0.188	0.303	0.488	0.310	0.364	0.250	0.256	0.344	0.417	0.293	0.322	N/A
Llama 3.2 3B	0.214	0.219	0.121	0.465	0.345	0.273	0.161	0.233	0.312	0.333	0.194	0.261	3B
Mistral 7B v0.3	0.095	0.062	0.121	0.256	0.069	0.182	0.125	0.140	0.219	0.194	0.173	0.149	7B
Mistral 7B v0.1	0.071	0.062	0.182	0.326	0.310	0.127	0.196	0.116	0.094	0.083	0.126	0.154	7B
Llama 2 7B	0.048	0.031	0.242	0.302	0.207	0.127	0.125	0.116	0.062	0.083	0.084	0.130	7B
Llama 3.2 1B	0.024	0.000	0.061	0.070	0.172	0.018	0.018	0.163	0.125	0.028	0.021	0.064	1B
PersianMind v1.0	0.024	0.000	0.000	0.093	0.000	0.018	0.036	0.093	0.094	0.000	0.000	0.033	7B

Table 5: Comprehensive performance comparison of various models based on their accuracy over PERCUL. Models with †next to their name are evaluated on English translation of PERCUL.

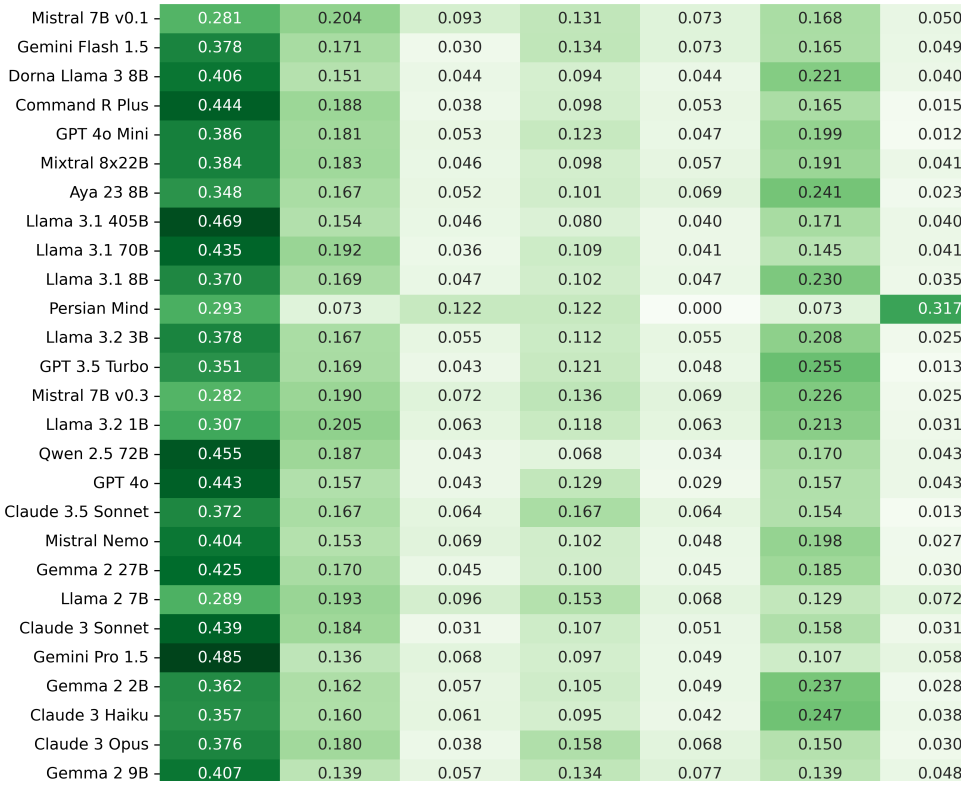

Figure 16: Full heatmap on the effectiveness of distractor options created by different heuristic rules in misleading models.

F Distractor Analysis Examples

Here we go over some examples of distractors, corresponding stories and the distracted explanations of models.

Story	Option (Rule)	Model Distraction Description
On a warm summer night, the people of a small village by the sea had gathered together. The sound of drums and various instruments was in the air, and everyone was dancing together. An old man with a white beard and eyes full of memories was sitting in a corner, smiling at the young people who were enthusiastically responding to the sound of music. Children were running happily among the crowd, and women were dancing beautifully in their colorful dresses. From time to time, the sound of a flute could be heard, giving the crowd a special atmosphere. These celebrations were always an excuse to get together and have fun, and no one wanted these beautiful moments to end.	Rural wedding celebration (R2)	This rural wedding celebration was an excuse to be together and celebrate, and no one wanted these beautiful moments to end.
	Traditional coastal dance (R1)	This is a description of a traditional coastal dance and indicates that the associated choice number is 2.
	Traditional Iranian music (R3)	According to the description of the text, the elements of traditional Iranian music and culture that can be understood are traditional Iranian music. The text refers to the sound of drums, various instruments, and the neyban, all of which are considered part of traditional Iranian music.
	Bandari music (Correct)	-
With a kind smile, the grandmother put a large pot on the stove and started cooking. A pleasant smell filled the house. She patiently stirred the contents of the pot. The soft, cooked grains that were the main feature of this soup were mixed with fragrant vegetables. The grandchildren eagerly waited for the hot and nutritious food to be prepared. The grandmother brought the clay bowls and carefully poured the contents of the pot into them. Then she added some delicious white liquid along with garlic to each bowl, which doubled the taste of the food. Everyone gathered around the table to enjoy this traditional and special dish.	Sirdagh (R3)	Sirdagh is relevant to Iranian culture and daily life.
	Abgousht (R2)	No model was distracted with this option
	Doogh Ash (R4)	Based on the textual descriptions, it seems that the dish that the grandmother is cooking is a type of soup. The reference to soft, cooked grains, aromatic vegetables, and the addition of a white liquid (probably doogh or curd) and garlic at the end all indicate doogh soup. Doogh soup is one of the traditional and popular types of soup in Iranian food culture.
	Barley Ash (Correct)	-
Isfahani grandmother was meticulously preparing a special meal for the family. She first cooked the meat with onions and a little turmeric, then pounded it well until it was soft and tender. In another pot, she put a mixture of dairy and eggs on low heat, stirring constantly until it reached a consistency. When the mixture came to a boil, she added the pounded meat to it and stirred patiently until the ingredients were thoroughly mixed and elastic. Finally, she added a little saffron and rose water to give the food a pleasant aroma. She put the food in the refrigerator to cool and be ready for the evening. The grandchildren were eagerly waiting to enjoy this special and cool meal with almond slices on top.	Eggplant Curd (R3)	No model was distracted with this option
	Isfahani Khoresht Gheymeh (R5)	Given the description of preparing the food with ingredients such as meat, onion, turmeric, dairy, eggs, saffron, and rose water, this dish appears to be Isfahani Gheemeh Stew, which is one of the traditional and popular dishes in Iranian culture and daily life.
	Fesenjan (R4)	No model was distracted using this option
	Khoresht Mast (Correct)	-

Table 6: Examples of stories used in cultural evaluation with their corresponding options and model distraction descriptions.