Advancing Persian LLM Evaluation

Sara Bourbour Hosseinbeigi[2], Behnam Rohani[1], Mostafa Masoudi[4], Mehrnoush Shamsfard[3], Zahra Saaberi[3], Mostafa Karimi Manesh[3], Mohammad Amin Abbasi[5],

[1]Sharif University of Technology, [2]Tarbiat Modares University,
[3]Shahid Beheshti University, [4]University of Tehran,
[5]Iran University of Science and Technology

Correspondence: s.bourbour@modares.ac.ir

Abstract

Evaluation of large language models (LLMs) in low-resource languages like Persian has received less attention than in high-resource languages like English. Existing evaluation approaches for Persian LLMs generally lack comprehensive frameworks, limiting their ability to assess models' performance over a wide range of tasks requiring considerable cultural and contextual knowledge, as well as a deeper understanding of Persian literature and style. This paper first aims to fill this gap by providing two new benchmarks, PeKA and PK-BETS, on topics such as history, literature, and cultural knowledge, as well as challenging the present state-of-the-art models' abilities in a variety of Persian language comprehension tasks. These datasets are meant to reduce data contamination while providing an accurate assessment of Persian LLMs. The second aim of this paper is the general evaluation of LLMs across the current Persian benchmarks to provide a comprehensive performance overview. By offering a structured evaluation methodology, we hope to promote the examination of LLMs in the Persian language.

1 Introduction

Large language models (LLMs) have become increasingly essential in natural language processing (NLP) across multiple domains, necessitating evaluation. While significant progress has been made in developing benchmarks and evaluation methods for high-resource languages such as English, the assessment of LLMs in low-resource languages like Persian remains limited.

The most advanced Persian language models are multilingual, transferring knowledge and skills from high-resource languages such as English to Persian. Although these models can grasp general concepts and skills that can be transferred from one language to another (Xu et al., 2023; Qi et al., 2023), they are unable to fully capture the complexity of Persian and learn knowledge accessible solely in Persian sources.

Despite the growth of multilingual LLMs, most studies have focused on worldwide languages, leaving Persian's particular linguistic, cultural, and contextual problems largely unexplored. Current benchmarks for Persian LLMs are often challenging a general skill or knowledge (Shariati et al.), which can be borrowed and generalized from an English setting to Persian.

Existing tools tend to focus on specific areas, such as basic NLP tasks (Khashabi et al., 2021), rather than providing a thorough evaluation of Persian-specific tasks. To address this issue, this study introduces new datasets designed specifically for evaluating LLMs in Persian. These datasets cover a wide variety of topics, including history, literature, social knowledge, and Persian cultural understanding. By utilizing these datasets, we can provide a more accurate and culturally relevant assessment of Persian LLMs.

Our primary goal is to evaluate the performance of open-source and proprietary LLMs using the available Persian benchmarks along with a translated version of ARC (Clark et al., 2018), and a set of newly created Persian benchmarks: PeKA and PK-BETS. These benchmarks are intended to test models on tasks that require not only general language skills, but also a thorough comprehension of Persian culture, linguistic intricacies, and contextual reasoning. The study examines LLMs using a variety of NLP tasks, including multiple-choice, short-answer, and open-ended questions. These tasks were carefully chosen to highlight the most challenging aspects of Persian language comprehension and generation. We examine the performance of the best models in Persian and how much they borrow from other high-resource languages by adding a wide range of domains and themes.

Along with benchmarking models over multiple-

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 2711–2727
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

choice questions, it is also highly important to evaluate the performance of the models on open-ended questions and compare their generative capabilities. We use LLM-as-a-Judge for this purpose, using two methods: pairwise comparison and single-answer grading (Zheng et al., 2023).

2 Related Work

In recent years, the evaluation of Large Language Models (LLMs) has become increasingly important, particularly for languages with limited resources ike Persian. Substantial progress has been made in building benchmarks and toolkits for evaluating LLMs in high-resource languages, but many issues within the low-resource languages remain unexplored.

2.1 Open-source widely-used English benchmarks

Open-source benchmarks like MMLU (Hendrycks et al., 2021), ARC (Clark et al., 2018), Winogrande (Sakaguchi et al., 2021), BBH (Suzgun et al., 2023), HellaSwag (Zellers et al., 2019), GSM8k (Cobbe et al., 2021), and TruthfulQA (Lin et al., 2022) are widely used to assess language models in diverse tasks. MMLU evaluates models across 57 academic and professional subjects, whereas ARC tests scientific reasoning through standardized exam questions. Winogrande assesses commonsense reasoning by resolving ambiguous pronouns, whereas BBH addresses challenging logic and mathematics challenges. HellaSwag is a dataset for physically situated commonsense reasoning. GSM8k tests models on grade-school-level math word problems, and TruthfulQA measures a model's truthfulness and resistance to generate misleading or inaccurate information. By translating several of these benchmarks into Persian, we ensure that the models' cross-lingual performance is thoroughly evaluated.

2.2 Open-source Persian benchmarks

To test LLMs in Persian, several datasets are available for different NLP tasks like NER (Poostchi et al., 2016; Shahshahani et al., 2018), NLI (Amirkhani et al., 2023; Khashabi et al., 2021), sentiment analysis (Khashabi et al., 2021; Sharami et al., 2020; Nazarizadeh et al., 2022), paraphrasing (Khashabi et al., 2021; Mohtaj et al., 2022; Sadeghi et al., 2022), irony detection (Golazizian et al., 2020), reading comprehension (Khashabi et al., 2021; Abadani et al., 2021; Ayoubi, 2021; Darvishi et al., 2023), summarization (Farahani et al., 2021b; Behmadi Moghaddas et al., 2013; Salemi et al., 2021; Farahani, 2020; Hasan et al., 2021), translation (Khashabi et al., 2021; Kashefi, 2020; Karimi et al., 2018), and spell correction (Persian, 2021; Mirzababaei et al., 2013). Recently, Persian-MMLU was introduced by Ghahroodi et al. (2024), covering substantial gaps in previous research by offering a culturally nuanced dataset specifically designed for Persian, with 20K multiple-choice questions across 38 diverse domains, ranging from elementary to secondary education levels. Also, Khashabi et al. (2021) provided a test set in multiple-choice question format with 2.4K instances derived from Persian educational texts, such as exams and employment tests in three domains: literature, commonsense, and mathematics. Abaskohi et al. (2024) also released two new benchmarks of math problems consisting of 279 samples drawn from elementary school questions and entrance exams for talented students.

2.3 Evaluation of Persian language

As multilingual LLMs have improved, demonstrating considerable promise for comprehending and producing Persian material, the necessity to assess their performance across several facets of the Persian language has grown. Abaskohi et al. (2024) took on this challenge by researching GPT-3.5 (Brown et al., 2020) and GPT-4 (OpenAI et al., 2024) as closed-source models, alongside OpenChat3.5 (Wang et al., 2024) as an open-source alternative. They evaluated the models across different NLP tasks, incorporating an investigation into various prompt engineering techniques to enhance performance. Khashabi et al. (2021) evaluated multilingual language models like mT5 (Xue et al., 2021) and Persian-specific models such as ParsBERT (Farahani et al., 2021a). They also offered a number of BERT-based models for various tasks that have been fine-tuned using Persian datasets. Additionally, Ghahroodi et al. (2024) provided a benchmark to assess the performance of both open-source and closed-source LLMs, such as GPT-4, XVERSE[1], Aya (Üstün et al., 2024), and Claude (Anthropic, 2024). There are also two GitHub projects dedicated to benchmarking Persian. MofidAI (Fallahnejad and Zarezade) benchmarks HuggingFace Persian language models, such as mT5 and BERT-based models, using a variety of

[1] https://github.com/xverse-ai

available datasets. The repository addresses challenges such as the absence of standardized evaluation codes and specified test sets. ParsBench (Shariati et al.) is a toolset designed exclusively to benchmark LLMs in Persian. It covers a wide range of tasks, including sentiment analysis, machine translation, and multiple-choice question answering, using datasets like ParsiNLU and Persian-MMLU. ParsBench has also developed a leaderboard on Huggingface that ranks several LLMs, including well-known open-source and closed-source models, as well as those fine-tuned specifically for Persian. Our work has broadened the evaluation of the Persian language from multiple perspectives. This includes establishing a generalizable assessment framework, expanding the variety of benchmarks available and examining a diverse range of LLMs.

3 Extending Persian benchmarks

We provide two benchmarks designed exclusively for the Persian language, which cover a wide range of domains and tasks. These benchmarks are developed with a thorough grasp of Persian culture, ensuring that the majority of examples are native and contextually relevant to real-world use cases within the Persian-speaking community. Our work fills a major gap in the availability of Persian-centric benchmarks for evaluating LLMs in Persian and identifying model shortcomings. Each benchmark is described separately below.

3.1 PeKA: Persian Knowledge Assessment

This dataset is constructed so that answering these questions requires knowledge about Persian community, particularly Iran, and its culture from a variety of perspectives. It is worth noting that the majority of the essential information is available on the internet. This data set contains 3600 multiple-choice questions divided into 12 different categories, each with 300 high-quality examples. The Categories are as follows: *history, literature, religion, general knowledge, geography, nature, art, music, television shows, movies, food, and sports* which cover a wide range of cultural and native topics for Persian speakers.

Data Construction Method. The data for PeKA was derived from *Quiz of Kings*, a popular quiz game created by an Iranian gaming firm. This game allows participants from across Iran to compete in a variety of knowledge areas. While the majority of the questions are publicly available through the game, additional curated data was obtained directly from the producers for research purposes. The questions were verified by specialists appointed by the company, ensuring their factual correctness and contextual relevance. Additionally, the game's terms of use, agreed upon by users at the start, allow for the use of user reactions to assess the quality of questions.

Each question is filtered through numerous stages to ensure that it is correct and valid. Initially, the same user base in the app votes to identify defective or incorrect questions. The chosen questions are then verified by specialists in each discipline. Each row of the data includes additional information on the number of people who chose each option or did not answer the question at all. The questions are then classified into three levels of difficulty: easy, medium, and hard based on the ratio of the people who have chosen the correct answer in each question (including those who have not chosen an answer within the limited time). (See Figure 1). Certain examples of dataset are illustrated in Appendix C.

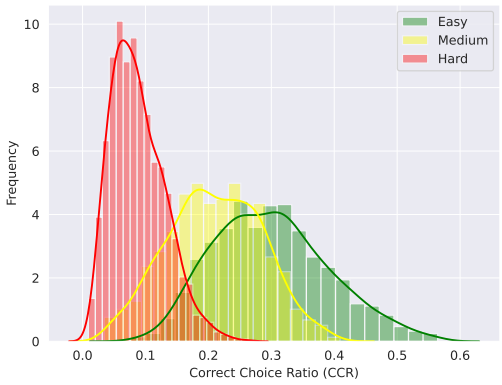

Figure 1: Distribution of correct choice ratio for each difficulty class

3.2 PK-BETS: Persian Knowledge - Bias Ethics Toxicity and Skills

This benchmark consists of a comprehensive set of domains meant to test the model's knowledge and language abilities in Persian. One major shortcoming in the research is the poor evaluation of models for Persian text generating tasks. To tackle this issue, PK-BETS is designed to include a variety of question types, such as multiple-choice, short answer, long answer, and open-ended questions. This allows us to evaluate not only the model's

capabilities using multiple-choice questions, but also its ability to generate coherent and meaningful content. The dataset is organized into 16 distinct categories so that each of them can fill a gap in the evaluation of the Persian language. Appendix D provides more information on these categories.

Data Construction Method The primary objective of this dataset is to identify and analyze the cultural and contextual weaknesses of the Llama 3 model, which, at the time, was the best-performing open-source model supporting Persian. The dataset creation process involved defining culturally specific categories related to Iran. For each category, we collaborated with one or more domain experts who were free to select reliable sources based on their expertise. These experts formulated questions directly from these sources, ensuring the cultural and contextual accuracy of the data.

The primary aim of PK-BETS is to identify areas where Llama-3 exhibited weaknesses in tasks specific to Persian cultural and contextual knowledge. Approximately 50% of the questions were selected where Llama-3 provided accurate answers, while the remaining 50% highlighted its hallucinations or errors.

When creating this dataset, we start by carefully selecting varied sources to extract questions that are culturally and contextually relevant to the Persian population. These sources include books, university tests, and expert insights. Questions are generated directly from these sources or by human authors. We then consider Llama-3 70B's answers to these questions. We filter and curate a subset of the dataset, maintaining a suitable proportion of the questions that the model answers properly and instances where the model hallucinates or fails to provide accurate responses. This organized method produces a rich and diverse dataset, showcasing the model's strengths and limitations in interpreting and addressing culturally relevant queries.

The ground truth for PK-BETS was established by the contributing experts independently of Llama-3's answers. These experts formulated questions and determined correct answers based on their chosen credible sources. Llama-3's outputs were used only to analyze its performance, and its responses did not influence the dataset's ground truth. This process ensures that the benchmarks derived from PK-BETS are robust, unbiased, and reflective of expert knowledge.

The dataset's questions were sourced and selected in collaboration with domain-specific specialists from multiple categories. A variety of sources were used for each topic to ensure diversity and richness in the dataset. All references were carefully vetted by specialists to ensure quality and relevance. Examples include:

Legal Based on the Islamic Republic of Iran's Constitution and other legal references regularly utilized by the Iranian judiciary.

Religious Based on widely used references, such as *Ayin-e Zindegi* (Life's Ethics), a university textbook, and other culturally significant religious texts.

Medical Sourced from books and university curricula frequently utilized in Iranian educational and medical contexts, covering both mainstream and traditional Iranian medicine.

Cultural and Social Inspired by Persian literature, proverbs, and history, with references to educational books and curated knowledge from Iranian scholars.

To guarantee authenticity and diversity, experts were allowed to create questions on their own and select trustworthy sources that were pertinent to their fields. These sources were either widely recognized or based on their professional expertise.

3.3 Translation of widely-used open English benchmarks

Numerous benchmarks have been developed to evaluate LLMs in English, covering a wide range of tasks and dimensions. To extend these evaluations to Persian, we translated the ARC dataset using GPT-4o and GPT-4o-mini model APIs. Our analysis examines the performance degradation of multilingual models when comparing the original English ARC dataset to its Persian translation (See Table 1). We ensured that the translation procedure for this specific benchmark was of high quality. Appendix B provides more information about the translation process and quality.

4 Benchmarking open-source LLMs for Persian

Several benchmarks are provided including the most useful available Persian datasets as well as newly created and translated benchmarks on various open-source and proprietary LLMs in different sizes. Our evaluation framework organizes and

	Params	ARC Challenge *0-shot*	
		Original	Translated
Llama 3	8B	77.59	62.23
	70B	92.36	86.00
Llama 3.1	8B	80.08	65.15
	70B	93.81	88.15
Aya-23	8B	66.69	60.08
	35B	83.43	75.27
Qwen 2	7B	83.94	65.92
	72B	94.16	87.46
Qwen 2.5	7B	86.35	68.49
	14B	91.41	82.23
	32B	92.10	86.00
Command-r	35B	83.17	72.87

Table 1: The drop in models' accuracy from the original ARC benchmark (challenging subset) to the translated version in Persian.

cleans our experiments (see Appendix E for more details), providing insight into the current status of large language models in Persian. This review aims to identify gaps and potential for improvement, guiding future research and development in the subject. The models, tasks, prompts, and required resources are detailed below.

Models. Some of the examined models explicitly state that they only support specific languages but include Persian in their capabilities, whereas others officially include Persian as part of their multilingual functionality. We conduct our evaluations on Llama3, Llama3.1 (Dubey et al., 2024), Qwen2 (Yang et al., 2024), Qwen2.5 (Team, 2024), Aya-23, and Command-r-v01[2] models. Table 2 summarizes the evaluated models, their parameter counts, and whether they officially support the Persian language. Also, models' setup during experiments is explained in Appendix A

Model Name	#Params	Persian supported
Meta-Llama-3.1-Instruct	8B,70B	No
Meta-Llama-3-Instruct	8B,70B	No
Qwen2-Instruct	7B,72B	Yes
Qwen2.5-Instruct	7B,14B,32B, 72B	Yes
c4ai-Aya-23	8B, 35B	Yes
c4ai-Command-r-v01	35B	Yes*

Table 2: Evaluated models and Persian support status. *: Persian data in just pre-training.

[2]https://huggingface.co/CohereForAI/c4ai-command-r-v01

Tasks. Our test data are picked from available open-source Persian datasets, our own new benchmarks, and translated datasets to ensure a thorough evaluation and proper ordering of LLMs based on their skills in Persian. We aim to represent a model's abilities in important tasks and domains. Assessed datasets, their size and tasks are shown in Table 3.

Dataset	Task	Test Size
ParsiNLU	MCQ	1k
	NLI	1.7K
	Query Paraphrasing	1.9K
PersianMMLU	MCQ	20K
ARC (translated)	MCQ	3.5K
PeKA	MCQ	3.6K
PK-BETS	MCQ + Open-ended Questions.	4K
PQuAD	Reading Comprehension	8K

Table 3: Datasets which are used for our evaluation tests. MCQ: multiple-choice question. NLI: Natural Language Inference (Textual Entailment).

Prompts. We run our experiments in different settings based on the dataset and test various prompt structures for each task to identify a prompt that achieves the highest score. Since our assessment is focused on the Persian language, we try to find the best prompts whose content is almost completely Persian, but English prompts are also tested. We explore the impact of using or not using the chat template, along with various general and task-specific system prompts, and proceed with experiments based on the best-performing setup.

Resources. We use 1, 2, or 3 Nvidia A100 80G to evaluate open-source models in the range of ~8, ~35, or ~70 billion parameters respectively.

5 Results

The evaluation results of open-source and proprietary models are shown in Table 4. When we compared Qwen and Llama models, we observed that Qwen struggled in knowledge-based questions, whereas Llama demonstrated a considerably deeper understanding and awareness of Persian history, culture, and almost every other domain. Llama 3.1 performed particularly well, excelling in the two new benchmarks provided: PeKA and PK-BETS.

However, when it comes to reasoning and textual entailment, Qwen models with around 7 billion parameters outperformed other models of the same size. Particularly, it was able to answer difficult math and logic questions from PersianMMLU.

Moreover, Qwen-2.5-32B demonstrated skills on par with the models of size ~70B and even performing better than them in several benchmarks despite having a lower capacity to reach them in knowledge-based benchmarks.

5.1 Analysis of PeKA based on categories

We conduct a detailed analysis of the performances of open-source models in different sizes on the PeKA benchmark, which focuses on Persian knowledge (See Table 5). A trend in model ranking is observed across the categories in this benchmark in Figure 2, showing the correlation between the knowledge capacity of models and their number of parameters. Furthermore, it is observed that Llama models' knowledge is strictly superior to that of Qwen's models when it comes to Persian in almost every category.

Generally, the models demonstrate stronger knowledge in more prominent and frequently covered areas, such as history, general knowledge, and religion. In contrast, their performance is weaker in less emphasized or specialized categories, such as TV, music, cinema, and food, where little to no information is available about them in other languages like English. This suggests that the models are better at understanding widely known and repeated information, carrying it from rich languages such as English to Persian, but struggle with less common or more culturally specific knowledge only available in Persian sources.

5.2 Analysis of PK-BETS categories

The performance of the models on different categories of PK-BETS (only multiple-choice questions) are presented in Table 6.

Cultural and Contextual Clarity In categories with legal or religious questions, the prompts explicitly instructed the models to provide answers aligned with Iranian cultural, legal, and medical norms. For instance:

Law Prompts emphasized alignment with Iranian law, including phrases like *"Answer this question based on Iranian's laws."*

Medicine Prompts specified whether they referred to traditional Iranian practices or modern medicine to reduce ambiguity.

Traditional Medicine Questions based on traditional Iranian medicine were clearly categorized, emphasizing their cultural significance and distinguishing them from conventional medical practices.

When confronted with questions in a Persian-specific context, existing multilingual, cutting-edge models perform badly in disciplines such as medicine and law. They lack essential linguistic skills in identifying metaphor and irony in Persian writing, lacking an understanding of toxicity, human bias and different emotions. These domains are where Persian and English style differs tremendously. This is where a knowledge or a skill, based solely on English writing, cannot be applied to the same writing in Persian, which is why it can hurt the performance of multilingual models in the above-mentioned tasks.

5.3 Evaluation by LLM-as-a-Judge

It is crucial to go beyond the typical evaluation of LLMs by benchmarks, which only focuses on multiple-choice questions. Therefore, we also assess and compare models on generation tasks that provide significant insights into their generative capabilities. The generation tasks included in the PK-BETS dataset offer a valuable resource for this type of evaluation. We evaluate smaller (~7-8B) open-source LLMs with two evaluation methods, (i) **pairwise comparison** and (ii) **single-answer grading** (Zheng et al., 2023). We also compare four open-source and proprietary models as judges for single-answer grading method.

5.3.1 Pairwise comparison of LLMs

In this method, the answers of two models are given to GPT-4o as a judge to decide if one model's response is superior, both are equally good (tie), or both are equally bad (bad tie). We selected a subset of 100 samples from the PK-BETS generation questions, ensuring that different categories and topics are used in the final evaluation. The pairwise comparison results for five different models are presented in Figure 3.

A subset of size 100 is picked for human evaluation. This is later used for measuring the agreement of judge's decision with human preferences using Cohen's kappa score (Cohen, 1960; McHugh, 2012). See table 7.

5.3.2 Single answer grading evaluation

To assess models' generating capabilities in Persian, we present a well prepared dataset of 40 questions. These questions ask the model to tell a story,

	Params	PeKA *0-shot*	PK-BETS (MCQA) *0-shot*	Khayyam Challenge *0-shot*	ParsiNLU (MCQA) *0-shot*	ParsiNLU (NLI)*	ParsiNLU (QQP)*	PQuAD *2-shot*
GPT-4o	N/A	**81.66**	**73.96**	**52.26**	–	–	–	–
GPT-4o-mini	N/A	62.90	64.68	–	–	–	–	–
Qwen 2.5	7B	37.59	45.94	38.68	41.52	70.71	83.56	71.94
	14B	42.62	52.84	41.87	48.00	78.24	84.97	78.27
	32B	47.68	58.82	49.19	53.80	**81.53**	86.63	81.31
	72B	54.20	61.30	51.20	57.52	81.35	87.73	79.06
Qwen 2	7B	37.32	47.81	37.01	41.71	69.28	82.67	51.34
	72B	52.57	60.65	47.91	53.05	77.76	82.51	–
Llama 3	8B	41.59	47.24	34.92	40.09	60.60	80.69	78.43
	70B	57.90	58.52	45.12	53.90	78.66	88.15	66.17
Llama 3.1	8B	44.10	48.76	35.70	42.67	59.41	80.58	77.37
	70B	62.38	64.12	48.65	**58.57**	78.54	**88.47**	**84.39**
Aya-23	8B	41.12	42.73	31.17	36.86	57.14	60.49	65.63
	35B	52.99	53.84	37.20	42.48	67.36	87.58	74.48
Command-r	35B	52.99	50.63	36.32	40.48	59.77	84.50	62.3

Table 4: Accuracy of open-source and proprietary models across various benchmarks. The best-performing model is highlighted in bold for each benchmark. *On NLI and QQP benchmarks, models are evaluated in different few-shot settings and the maximum score is considered. Maximum of 0,2,5,10-shots for QQP and maximum of 0,3,5,10-shots for NLI.

	Llama 3		Llama 3.1		Qwen 2		Qwen 2.5				Aya-23		Command-r	GPT-4o	GPT-4o mini
	8B	70B	8B	70B	7B	72B	7B	14B	32B	72B	8B	35B	35B	N/A	N/A
History	37.33	51.67	39.00	60.33	35.67	45.33	37.00	39.67	47.33	73.66	47.00	46.67	48.33	**90.66**	79.66
Literature	32.33	37.33	34.00	44.33	32.33	39.00	28.33	31.00	36.00	53.33	31.67	37.33	34.67	**81.66**	60.66
Religion	48.33	72.33	55.33	74.33	44.67	61.00	43.33	48.33	60.67	65	45.00	63.33	59.33	**89.333**	71.33
Knowledge	52.00	72.33	54.00	74.00	42.67	60.00	49.50	51.00	51.67	68.56	51.83	69.57	67.67	**90.30**	78.26
Geography	52.67	79.67	58.00	84.00	42.00	68.67	40.33	58.67	62.33	62.33	47.00	68.00	66.67	**90.66**	75.33
Nature	40.00	58.00	40.67	62.67	34.33	46.00	35.33	37.33	46.33	52.33	47.00	48.67	53.33	**86.0**	65.33
Art	36.33	48.00	36.00	51.00	34.33	50.67	34.33	36.00	43.67	64.66	41.67	47.33	47.33	**87.0**	69.0
Music	40.67	58.33	45.00	**65.67**	42.00	56.67	42.00	41.33	50.00	35	45.33	57.00	56.33	63.33	42.33
TV	27.00	31.00	29.67	32.67	27.00	26.67	21.00	21.00	21.67	24.33	24.00	29.00	29.00	**56.66**	36.0
Cinema	32.00	49.33	35.33	55.67	30.67	45.00	32.00	37.00	40.00	50	33.67	42.33	47.33	**82.0**	57.66
Food	56.19	70.90	59.20	76.59	49.50	71.91	49.50	54.85	61.54	44	51.84	42.33	69.23	**78.0**	53.66
Sports	40.67	58.33	45.00	65.67	42.67	56.67	42.00	41.33	50.00	57.33	45.33	57.00	56.33	**84.33**	65.66

Table 5: Performance (%) of various models across different categories of PeKA benchmark.

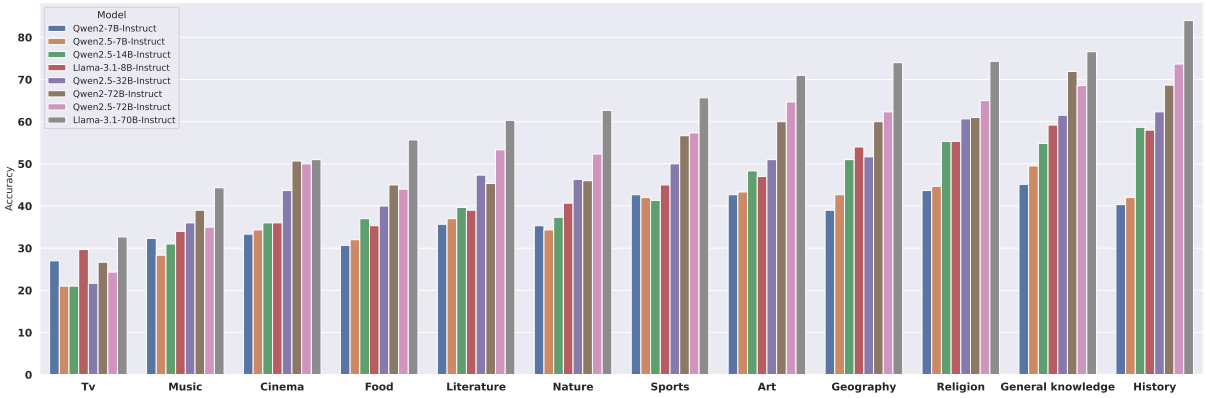

Figure 2: Llama vs Qwen; PeKA results for each category.

describe certain scenarios from a specific point of view, or characterize a conversation between two people with different viewpoints. The model's responses are then scored on a scale of 0 to 5 by

	Llama 3		Llama 3.1		Qwen 2		Qwen 2.5				Aya-23		command-r	GPT-4o	GPT-4o-mini
	8B	70B	8B	70B	7B	72B	7B	14B	32B	72B	8B	35B	35B	N/A	N/A
Persian Language	44.50	65.25	42.50	68.25	57.25	74.00	56.25	65.00	71.50	75.25	36.50	52.25	47.75	**83.0**	67.5
Respecting Others' Rights	40.00	60.00	41.67	60.00	45.00	65.00	41.67	41.67	68.33	73.33	50.00	65.00	56.67	**81.66**	68.33
Text Generation	50.91	61.82	43.64	69.09	60.00	63.64	41.82	49.09	67.27	65.45	60.00	63.64	52.73	**83.63**	76.36
Bias	45.29	44.12	41.76	63.53	35.29	65.29	41.18	54.12	58.24	57.64	36.47	42.35	45.29	**71.76**	61.17
Emotion Analysis	49.8	57.6	56.2	63.8	49.2	60.6	44.2	54.4	55.4	54.6	51.4	58.4	52.4	**69.8**	65.4
Medicine	37.50	52.50	38.12	55.62	40.00	49.38	35.62	45.00	46.88	54.37	38.75	40.00	42.50	**73.12**	59.37
Paraphrase	76.0	76.0	62.0	78.0	74.0	80.0	74.0	54.0	**86.0**	80	58.0	62.0	58.0	**86.0**	**86.0**
Recommendation	54.76	69.05	47.62	76.19	53.57	65.48	41.67	47.62	59.52	69.04	50.00	60.71	58.33	**77.38**	73.80
Style Transfer	62.0	60.0	76.0	66.0	46.0	64.0	64.0	54.0	70.0	74	38.0	68.0	62.0	**86.0**	74.0
Toxicity	30.97	**54.87**	49.56	**54.87**	34.51	32.74	44.25	44.25	42.48	48.67	27.43	47.79	52.21	45.13	40.70
Encyclopedic Knowledge	43.5	61.0	42.5	66.5	33.0	53.0	29.5	42.5	47.5	57.5	33.0	54.5	49.5	**84.5**	68.0
Law	48.10	52.38	48.10	60.00	48.10	52.86	45.24	50.00	51.90	53.33	43.33	48.57	48.10	**63.33**	55.23
Metaphor	28.33	38.33	36.67	56.67	40.00	48.33	28.33	36.67	51.67	51.66	25.00	40.00	30.00	**66.66**	51.66
Irony	56.0	60.0	55.0	65.0	54.0	56.0	56.0	51.0	64.0	58	50.0	62.0	55.0	65.0	**66.0**
Empathy Intimacy Trust	79.31	79.31	74.14	87.93	70.69	82.76	68.97	77.59	82.76	82.75	68.97	79.31	81.03	**91.37**	86.20
Religion	45.71	62.86	60.00	57.14	37.14	60.00	48.57	51.43	51.43	57.14	34.29	57.14	54.29	**80.0**	71.42

Table 6: Performance (%) of various models across different categories of PK-BETS benchmark (only multiple-choice question-types).

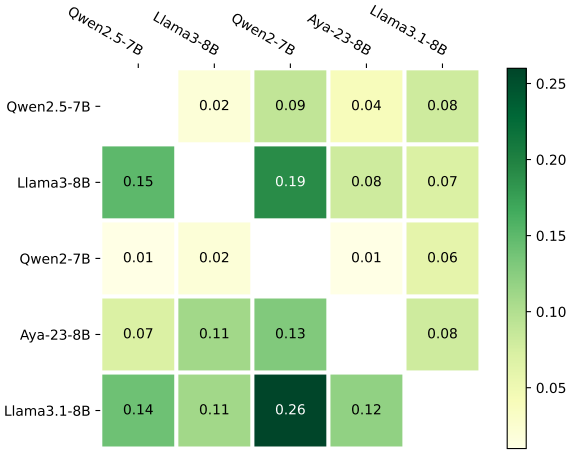

Figure 3: Win rate of models against each other in pairwise comparison method using GPT-4o as judge. For each pair of models, they both respond to 100 questions of generative problem-types in PK-BETS.

	GPT-4o	GPT-4o-mini	LLaMA
κ	0.54	0.41	0.24

Table 7: Cohen's κ score of different models as judge for the task of pairwise comparison with a given ground-truth.

GPT-4o for six linguistic criteria: coherency, fluency, grammar, naturalness, repetition, and diversity. The evaluation outcomes of various models' responses are provided in Table 8.

In general, the responses generated by Aya-23-8B showcased an extremely higher quality in terms of creativity, storytelling, and following natural linguistic structures than the other models. Aya-23-8B supports Persian (Table 2), which contributes to its outstanding performance in certain areas. Additionally, Llama 3.1, despite its substantial multilingual capabilities, is behind Qwen2 in terms of quality of generated text. However, Llama 3.1 significantly outperforms the previous Llama 3, demonstrating a notable improvement. The major problem with Llama 3 is that it tends to respond in English even though the input prompt is given in Persian, which radically lowers the language score in the final results.

	Llama-3	Llama-3.1	Qwen2	Qwen2.5	Aya-23
Coherency	2.17	4.1	4.3	4.35	**4.76**
Diversity	1.85	3.25	3.8	3.87	**4.2**
Fluency	2.02	3.72	4.02	4.1	**4.42**
Grammar	2.07	4.22	4.17	4.15	**4.42**
Naturalness	2	3.75	4.07	4.1	**4.47**
Repetition	1.65	2.7	3.07	3.275	**3.5**
Average	1.85*	3.63	3.92	3.99	**4.30**

Table 8: Performance comparison of candidate ~7-8B models across different LLM-as-a-Judge metrics to evaluate the quality of generated responses by models. *Low score of Llama 3 is due to sometimes generating responses purely in English when given a question in Persian.

Agreement of LLM judge with human evaluation We report Kendall's Tau correlation score, a metric used to measure the correlation between two sets of ordinal data (Kendall, 1938), between the scores assigned by the LLMs and four different human evaluators. For this purpose, a dataset is created consisting of 400 responses generated by 10 different models across 40 questions mentioned earlier. We then compute this measure for judge mod-

els Llama 3.1 70B, Qwen2 72B, GPT-4o mini, and GPT-4o, comparing their scores to human judgement, as shown in Table 9. A Kendall's Tau coefficient above 0.5 generally indicates a strong positive correlation. Both GPT-4o and GPT-4o mini showcase a higher alignment with human judgement compared to Llama 3.1 and Qwen2.

Criteria	LLaMA	Qwen2	GPT-4o mini	GPT-4o
Grammar	0.46	0.34	0.49	**0.54**
Fluency	0.47	0.51	0.51	**0.55**
Coherency	0.56	0.54	0.57	**0.60**
Naturalness	0.51	0.51	0.51	**0.56**
Diversity	0.60	0.62	**0.65**	0.62
Repetition	0.56	0.55	0.52	**0.60**

Table 9: Kendall's τ coefficient between LLM and human judges.

6 Conclusion

In this study, we addressed the significant gaps in evaluating large language models (LLMs) in Persian by introducing two new benchmarks. These benchmarks cover a diverse range of topics and tasks, providing a rich framework for assessing the proficiency of LLMs in Persian.

Through extensive experiments, we evaluated both open-source and proprietary models across various benchmarks, examining models of different sizes and capabilities. Our results show that Llama 3.1 and Qwen 2.5 outperformed others in the majority of tasks among open-source models. However, both models still lag behind the proprietary GPT-4o, which demonstrated a strictly superior performance in almost all domains and tasks. Additionally, in generation problem-types where LLMs act as evaluators, we found that GPT-4o exhibited better alignment with human judgment for Persian open-ended questions.

Interestingly, the smaller Aya-23-8B model generated responses that were well-suited to Persian linguistics, though it did not achieve the highest performance on the broader benchmarks. This suggests that while certain models may excel in specific linguistic tasks, their overall capacity to handle a wide range of challenges in Persian language remains limited.

Our findings highlight the need for ongoing development in proprietary and specially open-source models to fully capture the intricacies of the Persian language. Our benchmarks provide a critical foundation for future work in this area, enabling more accurate and culturally nuanced evaluation of LLMs in Persian.

7 Limitations

This research focuses primarily on Qwen, Llama, and CohereForAI's models, which proved capable in Persian language. However, it is worthwhile to also checkout other open-source models such as Mistral, Phi, and Yi. The other proprietary models such as Claude Sonet 3.5 should also be tested as it also demonstrated high-level capabilities in Persian.

We refrained from reporting the results in summarization and translation tasks because of poor quality and unreliability of current Persian benchmarks due to possible data contamination. It is helpful to introduce fresh and diverse datasets for translation and summarization tasks and report the corresponding results of current state-of-the-art models on them.

Although PK-BETS covers several gaps in the evaluation of models in Persian linguistic tasks, the size of this dataset is relatively small. Extending the current benchmark with more instances can be helpful.

We only test LLM-as-a-Judge for relatively smaller models due to the lack of computational resources and the high cost of proprietary models as evaluators.

There are still many aspects of the Persian language that are left unexplored. A further research into the performance of multilingual models on Persian-specific tasks and general tasks can enlighten strengths and weaknesses of these models.

8 Acknowledgement

We would like to sincerely thank National Artificial Intelligence Organization (NAIO) for providing the funds necessary to create and acquire the datasets, and the resources to evaluate the LLMs on these benchmarks.

References

Negin Abadani, Jamshid Mozafari, Afsaneh Fatemi, Mohamadali Nematbakhsh, and Arefeh Kazemi. 2021. Parsquad: Persian question answering dataset based on machine translation of squad 2.0. *International Journal of Web Research*, 4(1):34–46.

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.

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. The claude 3 model family: Opus, sonnet, haiku. Model Card Claude 3. Accessed: 2024-10-15.

Mohammad Yasin Ayoubi, Sajjad & Davoodeh. 2021. Persianqa: a dataset for persian question answering. https://github.com/SajjjadAyobi/PersianQA.

Behdad Behmadi Moghaddas, Mohsen Kahani, Seyyed Ahmad Toosi, Asef Pourmasoumi, and Ahmad Estiri. 2013. Pasokh: A standard corpus for the evaluation of persian text summarizers. In *ICCKE 2013*, pages 471–475.

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.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

ConfidentAI. Deepeval. https://github.com/confident-ai/deepeval. Accessed: 2024-09-27.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whit-

ney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zohreh Fallahnejad and Ali Zarezade. A benchmark for evaluation and comparison of various nlp tasks in persian language. `https://github.com/Mofid-AI/persian-nlp-benchmark`. Accessed: 2024-10-15.

Mehrdad Farahani. 2020. Summarization using bert2bert model on wikisummary dataset. https://github.com/m3hrdadfi/wiki-summary.

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

Mehrdad Farahani, Mohammad Gharachorloo, and Mohammad Manthouri. 2021b. Leveraging parsbert and pretrained mt5 for persian abstractive text summarization. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, pages 1–6.

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? In *First Conference on Language Modeling*.

Preni Golazizian, Behnam Sabeti, Seyed Arad Ashrafi Asli, Zahra Majdabadi, Omid Momenzadeh, et al. 2020. Irony detection in persian language: A transfer learning approach using emoji prediction. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2839–2845.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Akbar Karimi, Ebrahim Ansari, and Bahram Sadeghi Bigham. 2018. Extracting an English-Persian parallel corpus from comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Omid Kashefi. 2020. Mizan: A large persian-english parallel corpus. *Preprint*, arXiv:1801.02107.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

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.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

LinuxFoundation. Mlflow llm evaluation. `https://mlflow.org/docs/latest/llms/llm-evaluate/index.html`. Accessed: 2024-09-27.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3):276–282.

Behzad Mirzababaei, Heshaam Faili, and Nava Ehsan. 2013. Discourse-aware statistical machine translation as a context-sensitive spell checker. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 475–482, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Salar Mohtaj, Fatemeh Tavakkoli, and Habibollah Asghari. 2022. PerPaDa: A Persian paraphrase dataset based on implicit crowdsourcing data collection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5090–5096, Marseille, France. European Language Resources Association.

Ali Nazarizadeh, Touraj Banirostam, and Minoo Sayyadpour. 2022. Sentiment analysis of persian language: Review of algorithms, approaches and datasets. *Preprint*, arXiv:2212.06041.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li,

Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

PerSpellData: An Exhaustive Parallel Spell Dataset For Persian. 2021. Romina oji, nasrin taghizadeh and heshaam faili. In *Proceedings of The Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*, pages 8–14, Trento, Italy. Association for Computational Linguistics.

Hanieh Poostchi, Ehsan Zare Borzeshi, Mohammad Abdous, and Massimo Piccardi. 2016. PersoNER: Persian named-entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3381–3389, Osaka, Japan. The COLING 2016 Organizing Committee.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Reyhaneh Sadeghi, Hamed Karbasi, and Ahmad Akbari. 2022. Exappc: a large-scale persian paraphrase detection corpus. In *2022 8th International Conference on Web Research (ICWR)*, pages 168–175.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Alireza Salemi, Emad Kebriaei, Ghazal Neisi Minaei, and Azadeh Shakery. 2021. ARMAN: Pre-training with Semantically Selecting and Reordering of Sentences for Persian Abstractive Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9391–9407, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mahsa Sadat Shahshahani, Mahdi Mohseni, Azadeh Shakery, and Heshaam Faili. 2018. Peyma: A tagged corpus for persian named entities. *ArXiv*, abs/1801.09936.

Javad PourMostafa Roshan Sharami, Parsa Abbasi Sarabestani, and Seyed Abolghasem Mirroshandel. 2020. Deepsentipers: Novel deep learning models trained over proposed augmented persian sentiment corpus. *Preprint*, arXiv:2004.05328.

Shahriar Shariati, MohammadHossein barzrgari, and Masoud Marandi. Parsbench provides toolkits for benchmarking llms based on the persian language tasks. `https://github.com/ParsBench/ParsBench`. Accessed: 2024-10-15.

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.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024. Openchat: Advancing open-source language models with mixed-quality data. In *The Twelfth International Conference on Learning Representations*.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. 2024. Promptbench: A unified library for evaluation of large language models. *Preprint*, arXiv:2312.07910.

A Models' setup during inference

We fix the setting of models during inference so that experiments are comparable with each other. The specified values for these parameters are provided in Table 10

temperature	0.01
do_sample	True
top-k	50
top-p	1.0
repetition_penalty	1.0
max_new_tokens	based on task

Table 10: Hyperparameters of models during evaluation. All other parameters are the set to default values.

The max_new_tokens parameter that restricts the number of generated tokens varies based on the task; it is set to **32** for multiple choice questions where we just wish to extract the option's label or the actual answer, **680** for CoT configs that need space to reason and find the final answer (e.g. BBH), and also **512** for open-ended questions or long answers.

B Translation of widely-used open English benchmarks

We translate datasets with GPT-4o API from English to Persian. To choose the best model for translation, GPT-4, GPT-4o and GPT-4o-mini are compared. One hundred examples from the ARC dataset are chosen randomly and the responses of these models are compared. A native Persian speaker serves as the annotator, choosing the best-performing model for each example. Ultimately, GPT-4o, which demonstrated the best performance in terms of translation accuracy and fluency, was selected.

C PeKA Dataset Examples

Several examples of the PeKA dataset are presented in Table 11.

D PK-BETS Dataset Details

The dataset's 16 categories are explained in 10 different classes:

0. **Persian Language:** This section is designed to assess the proficiency of LLMs in various linguistic skills, including syntax, sentence structures, idioms, proverbs, and lexical semantics (such as analogies).
0. **Style transfer:** This assignment demands the model to interpret and adapt writings in a variety of styles, such as polite, colloquial, formal, and comical.
0. **Social Knowledge Understanding:** In this section, various aspects of social knowledge, such as emotion analysis, deciphering metaphors and irony, recognition of insults and aggression, as well as humanistic concepts like empathy, intimacy, and trust, are evaluated.
0. **Ethics and Bias:** This section evaluates the model for potential biases in several categories such as gender, religion, politics, and

Example	Label	Category
نام پسر کورش کبیر کدام است؟ ۱) اسفندیار ۲) کمبوجیه ۳) الب ارسلان ۴) اهورا ***English equivalent**: What is the name of Cyrus the Great's son?"* *1) Esfandiar 2) Cambyses 3) Alp Arslan 4) Ahura*	2	History
بام سبز از مناطق دیدنی کدام استان است؟ ۱) خراسان رضوی ۲) گلستان ۳) اصفهان ۴) گیلان ***English equivalent**: Bame Sabz is one of the scenic areas of which province?* *1) Khorasan Razavi 2) Golestan 3) Isfahan 4) Gilan*	4	Nature

Table 11: PeKA benchmark examples.

race. It also requires assessing the model's adherence to ethical principles and respect for the rights of others.

0. **Medicine:** This area includes general, specialized, and medical emergency knowledge in the fields of traditional medicine, such as medical terminology and drug names. These medical terms are typically provided in Persian and in some cases, English equivalent terms are provided.

0. **Law:** The questions are designed based on the officially approved Constitution of Iran, as well as the country's legal and criminal laws.

0. **Religion:** The religious questions are derived from the teachings of Islam, sourced from treatises written by religious scholars, Islamic texts, or frequently asked questions by the general population.

0. **Encyclopedic knowledge:** The questions are prepared in a variety of subjects, including social sciences, humanities, literature, sports, art, politics, economics, and culture. They are designed in such a way that the responses remain consistent and do not alter with time.

0. **Recommendation:** This component of the dataset is intended to evaluate the model's ability to make logical, accurate, and relevant suggestions. Each question describes a unique scenario in detail, allowing the model to make decisions and make recommendations based on the circumstances. Answering these issues usually involves a combination of common sense and the ability to use reasoning abilities successfully.

0. **Text Generation:** This section assesses the model's text generation capabilities across various tasks, which include: (i) generating a cohesive text on a topic that requires general knowledge across different areas, (ii) writing from a specific point of view, demonstrating an understanding of that perspective and its unique conditions, (iii) producing text in a particular speaking style, requiring familiarity with different tones and linguistic styles, (iv) continuing an incomplete text, (v) constructing sentences using a limited set of words, (vi) generating poems and songs, and (vii) paraphrasing text on different topics.

Several examples of dataset are presented in Table 12.

E Our Evaluation Framework

Evaluating LLMs in large-scale and systematic ways has many complexities that require a comprehensive framework. There are many open-source frameworks with specific features that help us evaluate open-source and API-based models on various benchmarks.

The Promptbench (Zhu et al., 2024) is one of the most popular frameworks that support a variety of models, datasets, prompt engineering techniques and attack methods, and other useful features to make it a unified and complete framework for LLM evaluation. Also, there are other tools such as DeepEval (ConfidentAI) and MLflow LLM Eval (LinuxFoundation) that can be used for the same purpose.

Inspired by open-source tools, we implement a framework to automate our evaluation experiments on various Persian benchmarks using customizable prompts. The framework is designed to be flexible and extendable, allowing users to easily add new datasets, integrate models, and actually create new

Example	Category
بهترین زمان تزریق واکسن آنفلوانزا چه موقعی است؟ ***English equivalent**: When is the best time to inject the flu vaccine?*	Medicine
آیا ارتکاب جرم باعث می‌شود تا ابد سوسابقه داشته باشیم؟ ***English equivalent**: Does committing a crime cause us to have a criminal record forever?*	Law
اکبری که یکی از مدیران ارشد شرکت است، امسال برای سومین سال پیاپی به عنوان مدیر نمونه انتخاب شد و از او تقدیر شد. اکبری خانم است یا آقا؟ دلیل انتخاب خود را توضیح بده. ***English equivalent**: Akbari, who is one of the senior managers of the company, was chosen as an exemplary manager for the third year in a row and was honored. Is Akbari a lady or a gentleman? Explain the reason for your choice.*	Bias
ماه مرا دیدی؟ چشمهایش میخندد و نفسش بوی بهار میدهد. چند استعاره و چه استعاره هایی در این متن وجود دارد؟ ***English equivalent**: How many metaphors and which type of metaphors are used in the following text? "Did you see my moon? Her eyes are smiling and her breath smells like spring."*	Persian Language

Table 12: Examples from some of PK-BETS dataset categories

evaluation pipelines. Our framework is currently under development to incorporate additional features and capabilities. In the following, the main components and capabilities of the framework are briefly explained.

5.1 Framework components

Our evaluation framework consists of four key components forming an evaluation pipeline.

Dataset. A variety of benchmarks, covering a wide range of NLP tasks, are supported, comprised of translated English datasets as well as new, culturally relevant Persian benchmarks, with the capability to add and handle custom datasets. Additionally, we supply both multiple-choice and open-ended benchmarks for assessing LLMs' understanding and generation capabilities. This component simplifies dataset integration, allowing easy loading and answer extraction, which is crucial for evaluation on a MCQA benchmark.

Model. This component supports both API-based models (OpenAI VLLM) and open-source models (Huggingface library), handling a range of decoder-only models like Llama, Qwen, and others. Models can be loaded in base (without chat-template) or instruct (with chat-template and optional system prompt) configurations, depending on the task. Model loading, default or extended tokenizer configuration, and control of model predictions are handled here.

Prompt. Accurate evaluation depends heavily on prompt engineering. Our system supports zero-shot, few-shot, and chain-of-thought prompting to leverage models' in-context learning and reasoning capabilities. Different datasets may require specific prompts, and our framework provides pre-designed prompts for each dataset to ensure accurate testing. A general prompt format is implemented to allow customization for various dataset needs.

Additionally, this component manages inference batching with two methods:

Simple batching is the default approach where data is processed in fixed batch sizes, set at the start of an experiment. It requires manual configuration based on available resources.

Smart batching is an additional, more dynamic method that adapts batch sizes based on available system resources. It monitors VRAM across CUDA devices, estimates memory usage per input, and adjusts batch sizes to maximize GPU utilization efficiently. This ensures optimal performance, especially in resource-constrained or variable environments.

Metric. We cover standard metrics such as accuracy, ROUGE, BLEU, and more, tailored for specific tasks. For generation tasks, traditional metrics are insufficient, so we introduce an LLM-as-a-Judge approach to evaluate the fluency, coherency, relevance, and correctness of responses. Details of LLM-as-a-Judge is described in Section 5.2.

5.2 LLM-as-a-Judge

Due to the limitations of metrics for generation tasks, we use the LLM-as-a-Judge evaluation method for three specific tasks:

(i) **Similarity Score:** This task involves scoring two texts based on several criteria, including meaning, sentiment, recall, precision, agreement, and an overall general assessment.

(ii) **Language Score (single-answer grading):** The evaluation focuses on the quality of the answer in terms of grammar, fluency, diversity, repetition, coherence, naturalness, and an overall general as-

sessment.

(iii) **Battle Score (pairwise comparison):** This involves comparing the answers generated by two models to a question against the ground truth to determine which model performs better.

The models used for judging must be compatible with the OpenAI API, such as GPT-4o or the Llama-3.1-70B model, both of which are accessible through the API.