Persian in a Court: Benchmarking VLMs In Persian Multi-Modal Tasks

Farhan Farsi
Amirkabir University of Technology
farhan1379@aut.ac.ir

Shahriar Shariati Motlagh
University of Mazandaran
s.shariati21@umail.umz.ac.ir

Shayan Bali
King's College London
shayan.bali@kcl.ac.uk

Sadra Sabouri
University of Southern California
sabourih@usc.edu

Saeedeh Momtazi
Amirkabir University of Technology
momtazi@aut.ac.ir

Abstract

This study introduces a novel framework for evaluating Large Language Models (LLMs) and Vision-Language Models (VLMs) in Persian, a low-resource language. We develop comprehensive datasets to assess reasoning, linguistic understanding, and multimodal capabilities. Our datasets include Persian-OCR-QA for optical character recognition, Persian-VQA for visual question answering, Persian world-image puzzle for multimodal integration, Visual-Abstraction-Reasoning for abstract reasoning, and Iran-places for visual knowledge of Iranian figures and locations. We evaluate models like GPT-4o, Claude 3.5 Sonnet, and Llama 3.2 90B Vision, revealing their strengths and weaknesses in processing Persian. This research contributes to inclusive language processing by addressing the unique challenges of low-resource language evaluation. Additionally, we release samples of our dataset to support further research in Persian multi-modal tasks[1].

1 Introduction

Large Language Models (LLMs) have undergone rapid advancements in recent years, particularly in multimodal frameworks (Zhang et al., 2024; Wu et al., 2023) that integrate and process diverse data types such as text, audio, and images. These breakthroughs have expanded the applications of LLMs across various domains, from conversational AI to content generation (He et al., 2024) and knowledge retrieval (Long et al., 2024). Multimodal LLMs demonstrate remarkable capabilities in aligning and interpreting visual-textual information (Ataallah et al., 2024), making them powerful tools for tasks that span different modalities (Nguyen et al., 2023). However, as the capabilities of LLMs grow, so does the need for rigorous evaluation methods to measure their effectiveness and ensure their outputs align with the intended goals (Huang and Zhang, 2024). Evaluating LLMs became a crucial area of research, especially when considering other languages rather than high-resource ones, where resources are abundant (Chang et al., 2023).

Although Persian is the native or second language for around 130 million people, high-quality datasets and benchmarks for the language remain limited (Agic et al., 2016). While researchers have introduced foundational datasets for pretraining LLMs (Sabouri et al., 2022; Salmasi and Kabir, 2023; Farsi et al., 2024), these resources are often single-modality. Evaluating LLM capabilities, such as reasoning, verbal intelligence, and multimodal reasoning, remains underexplored. Multimodal frameworks compound this challenge by requiring datasets that effectively pair text with complementary modalities, such as images, videos, while maintaining linguistic nuances (Hedderich et al., 2020).

In this study, we address these challenges by creating a comprehensive dataset designed explicitly for the Persian language. These datasets are constructed from scratch and designed to evaluate LLMs on multiple dimensions, including reasoning and verbal intelligence similar to prior works (Fu et al., 2024). Furthermore, we assess a set of large language models' performance with our framework, which measures the relative difficulty of different datasets and ensures a uniform evaluation across tasks (Li et al., 2023b). By investigating the reasoning capabilities of LLMs and their ability to interact with Persian linguistic constructs and multimodal data, we aim to uncover the extent of their knowledge base and adaptability to low-resource

[1] https://huggingface.co/AUT-NLP

Proceedings of the First Workshop of Evaluation of Multi-Modal Generation, pages 52–56
January 20, 2025.

languages. To tackle these challenges, this study proposes a multimodal evaluation framework focusing on creating datasets tailored to the Persian language and evaluating the capabilities of LLMs and Vision-Language Models (VLMs) in processing Persian. The framework assesses models across key dimensions, including reasoning abilities and visual-textual comprehension, while considering the difficulty of the datasets (Zhu et al., 2023). This work aims to evaluate the performance of current LLMs in multimodal contexts, providing a comprehensive assessment of their capabilities in Persian. This research fills critical gaps in low-resource language evaluation while contributing to developing inclusive, adaptable, and effective language processing models for diverse applications.

2 Related Work

The work related to our study can be divided into three main areas, discussed in the following sections. Together, these areas offer a foundational understanding of the challenges and opportunities in developing and evaluating models in multimodal and low-resource language settings.

Benchmarking LLMs and Dataset Development for Low-Resource Languages. Multimodal evaluation frameworks are critical for assessing models integrating and processing diverse data modalities. MME introduces a comprehensive benchmark for multimodal language models (MLLMs), evaluating perception and cognition abilities across 14 subtasks, enabling comparisons among advanced MLLMs (Fu et al., 2024). Similarly, SEED-Bench-2 categorizes MLLM capabilities hierarchically, incorporating tasks like image generation and providing detailed insights into model strengths and weaknesses (Li et al., 2023a).

However, developing benchmarks and datasets for low-resource languages like Persian is challenging due to limited resources and linguistic diversity (Sabouri et al., 2022). Multilingual benchmarks, such as IGLUE (Bugliarello et al., 2022), support zero-shot and few-shot learning across 20 languages, highlighting the potential of multilingual datasets but often lacking specific resources for Persian. While comprehensive benchmarks like GAOKAO (Zhang et al., 2023a) showcase LLMs' strengths in objective tasks, they also expose limitations in domain-specific challenges. For Persian, benchmarking efforts remained scarce, underlining the need for evaluation frameworks that reflect its unique linguistic and cultural features.

Benchmarks and Visual Reasoning for Vision-Language Models. Vision-Language Models (VLMs) are evaluated using benchmarks designed to test their ability to handle both visual and textual inputs (Xu et al., 2024). Benchmarks like VisIT-Bench focus on tasks such as accessibility assessments and image-caption generation (Bitton et al., 2023), while GEM evaluates multilingual vision-language tasks, including image and video interactions (Su et al., 2021). Visual reasoning benchmarks like GRASP test language grounding and intuitive physics understanding in video-based tasks (Jassim et al., 2023), while Multimodal-CoT uses chain-of-thought prompting to improve structured reasoning (Zhang et al., 2023b). Together, these benchmarks comprehensively evaluate VLM capabilities across diverse tasks. For dataset creation, we studied these works to establish best practices.

3 Datasets

We focused on investigating the evaluation of the multimodality attribute in Large Language Models (LLMs). Multimodal datasets contain data from multiple modalities, such as text, images, audio, video, or other structured/unstructured data types. For this study, we prioritized text-image data, enabling in-depth exploration of the model's ability to process and reason across these two modalities.

To create a comprehensive benchmark for VLMs in the Persian language, particularly focusing on multimodal varieties, we emphasized several key aspects like their reasoning skills, creativity, familiarity with linguistics in images, and knowledge about places in Iran. Our dataset has five distinct sets that we describe in the following paragraphs.

Persian-VQA: To create a VQA (Visual Question Answering) dataset in Persian, we used the Zhang et al. (2016) dataset, which is one of the most popular VQA datasets in English. This dataset contains 7,764 yes/no questions derived from 1,023 images. We translated the entire dataset into Persian using the GPT-4o model. To ensure the quality of the translated questions, we conducted a manual review of the generated dataset. An example of a record of this dataset is shown in Figure 1.

Persian-OCR-QA: Nowadays, OCR (Optical Character Recognition) has become one of the most important tasks due to its numerous applications (Peng et al., 2013; Singh et al., 2012). To

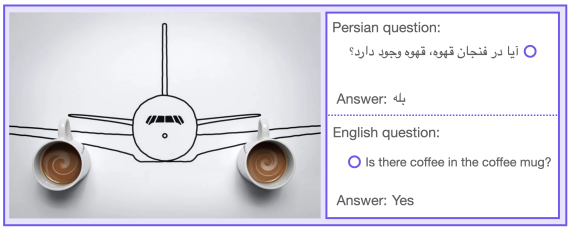

Figure 1: An example of Persian-VQA. Persian question-answer pair and its original English version.

evaluate the performance of current models on this task, we introduced new datasets to measure the performance of LLMs on OCR tasks in the Persian language. We used the Persian-OCR dataset, which contains 7,000 pages. Using GPT-4o-mini to make a question from the text and answer, we extracted ten question-answer pairs from each page, resulting in a comprehensive dataset of 70,000 entries.

Persian-VAR: To evaluate Vision Language Models (VLMs) in the domain of abstract reasoning, we introduce a novel dataset, Persian-VAR (Persian Visual-Abstraction-Reasoning), comprising 120 samples, inspired by Raven's Progressive Matrices (Carpenter et al., 1990). This non-verbal test is typically used to assess general human intelligence and abstract reasoning, and it serves as a non-verbal estimate of fluid intelligence. It is one of the most commonly administered tests to groups and individuals, from young children to the elderly. To create this dataset, we collected entrance exams for gifted middle and high schools in Iran, as illustrated in figure 2, providing a rich source of complex visual-abstraction-reasoning challenges that align with the cognitive capabilities assessed by Raven's matrices.

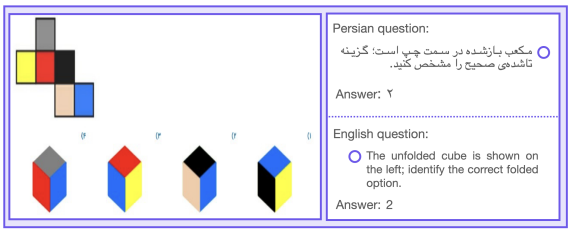

Figure 2: An Example of Persian-VAR. Persian question-answer pair and its original English version.

Persian-WIP: The Persian Word-Image Puzzle dataset assessed multimodal models' ability to integrate and process visual and textual information. By challenging models to combine visual cues with linguistic interpretation, this dataset evaluates their capability to manage complex inputs. Such tasks demand creative thinking and language skills, making it a robust framework for testing image recognition and language comprehension skills. This serves as both an educational tool and a benchmark for evaluating the effectiveness of multimodal systems. The dataset was compiled using crowdsourcing, crawling social apps like Telegram and Instagram, and generating images with AI models like Midjourney. Figure 3 displays a sample instance from the dataset.

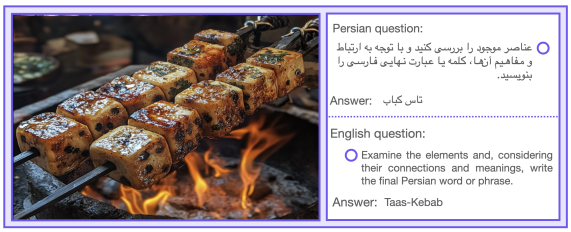

Figure 3: An Example of Persian-WIP. Taas-Kebab, a traditional Persian dish. The name combines "Taas" (dice) and "kebab" (grilled dish), referring to a dish made with diced kebab. When an image shows diced kebab, it represents Taas-Kebab in Persian.

Iran-Places: This dataset is designed to evaluate models on their knowledge of notable places in Iran, akin to the Persian version of (Weyand et al., 2020). It consists of over 500 images, with each province in Iran represented by at least seven images. This comprehensive coverage ensures a diverse representation of the country's geographical and cultural landmarks. An example of this is illustrated in Figure 4.

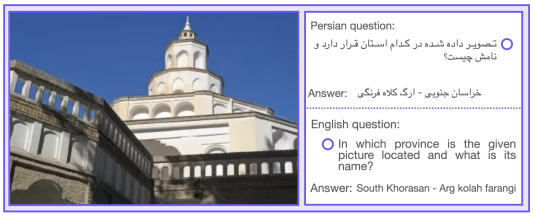

Figure 4: An Example of Iran-Places: Persian question-answer pair and its original English version.

4 Experiments

We tested current LLMs, such as ChatGPT-4o, Claude 3.5, and Llama 3.2, performance on the proposed benchmark. GPT-4o demonstrated superior performance in tasks requiring advanced reasoning and visual comprehension in Persian, indicating higher overall scores in cognitive tasks (Table 1). Claude 3.5 Sonnet excelled in text-based tasks like

Task	Metric	GPT-4o[a]	Claude 3.5 Sonnet[b]	Llama 3.2 90B Vision
P-VAR	Accuracy (%)	**16.22**	11.71	13.51
Persian-OCR-QA	BLEU-1 (%)	52.61	**57.53**	23.09
	ROUGE-L[c](%)	63.41	**77.47**	44.96
Persian-VQA	Accuracy (%)	**89.17**	85.86	82.89
	F1 Score (%)	**91.54**	87.82	86.78
Iran-Places	Relaxed Exact match[d](%)	16.44	**17.07**	16.43

[a] The GPT-4o-2024-09-03 version is used in this benchmark.
[b] The Claude-3.5-Sonnet-2024-10-22 version is used in this benchmark.
[c] We used the F1 score for ROUGE-L.
[d] We awarded 0.5 points if the name of the province or place was predicted correctly, 1 point if both were correct, and 0 points otherwise.

Table 1: Top 3 VLMs Performance on Different Multimodal Persian Tasks

OCR and text generation, suggesting strong textual processing capabilities. Llama 3.2 90B Vision showed balanced performance but with lower overall scores than the others.

All models struggled significantly with multimodal integration tasks, failing to achieve exact matches in the Persian Word-Image Puzzle, negatively impacting their overall multimodal scores. Similarly, low performance on tasks involving specific Iranian locations revealed limitations in culturally specific visual knowledge, affecting overall effectiveness in these areas.

These varied results, as detailed in Table 1, highlight the complexities of evaluating language models in Persian, showing strengths in specific areas but deficiencies in multimodal and culturally specific tasks. This underscores the need for further research and improved datasets to enhance model performance across diverse tasks.

5 Future Work

Our findings highlight the need for more specialized datasets for low-resource languages to improve model evaluation and performance. Future research should focus on developing new evaluation metrics, expanding multimodal datasets to include additional modalities like video and audio, and advancing model capabilities in handling complex multimodal tasks for the Persian language.

6 Conclusion

In this study, we introduced a framework for evaluating Large Language Models (LLMs) and Vision-Language Models (VLMs) in Persian, focusing on five specialized datasets: Persian-OCR-QA, Persian-VQA, Persian Word-Image Puzzle (P-WIP), Persian Visual-Abstraction-Reasoning (P-VAR), and Iran-Places. Our evaluations of GPT-4o, Claude 3.5 Sonnet, and Llama 3.2 90B Vision provided significant insights. GPT-4o excelled in abstract reasoning and visual question answering, highlighting its strong visual-linguistic integration capability. Claude 3.5 Sonnet showed superior performance in Persian-specific OCR tasks. Although all models performed similarly in geographical knowledge, they struggled with the Persian Word-Image Puzzle, revealing challenges in tasks needing creative multimodal synthesis.

References

Zeljko Agic, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. 2024. Minigpt4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. *arXiv preprint arXiv:2404.03413*.

Yonatan Bitton et al. 2023. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use. *ArXiv*.

Emanuele Bugliarello et al. 2022. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. *ArXiv*.

Patricia A Carpenter, Marcel A Just, and Peter Shell. 1990. What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. *Psychological review*, 97(3):404.

Yu-Chu Chang, Xu Wang, Jindong Wang, Yuanyi Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi,

Cunxiang Wang, Yidong Wang, Weirong Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qian Yang, and Xingxu Xie. 2023. A survey on evaluation of large language models. *ArXiv*, abs/2307.03109.

Farhan Farsi, Sadra Sabouri, Kian Kashfipour, Soroush Gooran, Hossein Sameti, and Ehsaneddin Asgari. 2024. Syntran-fa: Generating comprehensive answers for farsi qa pairs via syntactic transformation.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Yingqing He, Zhaoyang Liu, Jingye Chen, Zeyue Tian, Hongyu Liu, Xiaowei Chi, Runtao Liu, Ruibin Yuan, Yazhou Xing, Wenhai Wang, et al. 2024. Llms meet multimodal generation and editing: A survey. *arXiv preprint arXiv:2405.19334*.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strotgen, and D. Klakow. 2020. A survey on recent approaches for natural language processing in low-resource scenarios. pages 2545–2568.

Jiaxing Huang and Jingyi Zhang. 2024. A survey on evaluation of multimodal large language models. *arXiv preprint arXiv:2408.15769*.

Serwan Jassim et al. 2023. Grasp: A novel benchmark for evaluating language grounding and situated physics understanding in multimodal language models. *ArXiv*.

Bohao Li et al. 2023a. Seed-bench-2: Benchmarking multimodal large language models. *ArXiv*.

Zejun Li, Ye Wang, Mengfei Du, Qingwen Liu, Binhao Wu, Jiwen Zhang, Chengxing Zhou, Zhihao Fan, Jie Fu, Jingjing Chen, Xuanjing Huang, and Zhongyu Wei. 2023b. Reform-eval: Evaluating large vision language models via unified re-formulation of task-oriented benchmarks. *ArXiv*, abs/2310.02569.

Xinwei Long, Jiali Zeng, Fandong Meng, Zhiyuan Ma, Kaiyan Zhang, Bowen Zhou, and Jie Zhou. 2024. Generative multi-modal knowledge retrieval with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18733–18741.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq R. Joty, and Lidong Bing. 2023. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. *ArXiv*, abs/2306.11372.

Xujun Peng, Huaigu Cao, Srirangaraj Setlur, Venu Govindaraju, and Prem Natarajan. 2013. Multilingual ocr research and applications: an overview. In *Proceedings of the 4th International Workshop on Multilingual OCR*, pages 1–8.

Sadra Sabouri, Elnaz Rahmati, Soroush Gooran, and Hossein Sameti. 2022. naab: A ready-to-use plug-and-play corpus for farsi. *arXiv preprint arXiv:2208.13486*.

Ali Salmasi and Ehsanollah Kabir. 2023. Farsi text in scene: A new dataset. In *2023 13th International Conference on Computer and Knowledge Engineering (ICCKE)*, pages 510–514. IEEE.

Amarjot Singh, Ketan Bacchuwar, and Akshay Bhasin. 2012. A survey of ocr applications. *International Journal of Machine Learning and Computing*, 2(3):314.

Lin Su et al. 2021. Gem: A general evaluation benchmark for multimodal tasks. *ArXiv*.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2024. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5014–5022.

Xiaotian Zhang, Chun-yan Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023a. Evaluating the performance of large language models on gaokao benchmark. *ArXiv*.

Zhuosheng Zhang et al. 2023b. Multimodal chain-of-thought reasoning in language models. *ArXiv*.

Mingwei Zhu, Leigang Sha, Yu Shu, Kangjia Zhao, Tiancheng Zhao, and Jianwei Yin. 2023. Benchmarking sequential visual input reasoning and prediction in multimodal large language models. *ArXiv*, abs/2310.13473.