MEENA (PersianMMMU): Multimodal-Multilingual Educational Exams for N-level Assessment

Omid Ghahroodi[△], **Arshia Hemmat**[♠*], **Marzia Nouri**[♣*], **Seyed Mohammad Hadi Hosseini**[◇*], **Doratossadat Dastgheib**[△*], **Mohammad Vali Sanian**[◇†], **Alireza Sahebi**[◇†], **Reihaneh Zohrabi**[◇†], **Mohammad Hossein Rohban**[◇‡], **Ehsaneddin Asgari**[△‡], **Mahdieh Soleymani Baghshah**[◇‡]

[◇] Computer Engineering Department, Sharif University of Technology, Iran
[△]Qatar Computing Research Institute, Qatar
[♠] Computer Engineering Department, University of Isfahan, Iran
[♣] Independent Researcher

{oghahroodi98, arshiahemmat6, nouri.marzia.1999, mvs2667}@gmail.com
d_dastgheib@sbu.ac.ir, easgari@hbku.edu.qa
{hadi.hosseini17, alireza.sahebi, zohrabi, soleymani, rohban}@sharif.edu

Abstract

Recent advancements in large vision-language models (VLMs) have primarily focused on English, with limited attention given to other languages. To address this gap, we introduce MEENA (also known as PersianMMMU), the first dataset designed to evaluate Persian VLMs across scientific, reasoning, and human-level understanding tasks. Our dataset comprises approximately 7,500 Persian and 3,000 English questions, covering a wide range of topics such as reasoning, mathematics, physics, diagrams, charts, and Persian art and literature. Key features of MEENA include: (1) diverse subject coverage spanning various educational levels, from primary to upper secondary school, (2) rich metadata, including difficulty levels and descriptive answers, (3) original Persian data that preserves cultural nuances, (4) a bilingual structure to assess cross-linguistic performance, and (5) a series of diverse experiments assessing various capabilities, including overall performance, the model's ability to attend to images, and its tendency to generate hallucinations. We hope this benchmark contributes to enhancing VLM capabilities beyond English.

1 Introduction

In recent years, vision-language models (VLMs) (Radford et al., 2021) have rapidly advanced, driving breakthroughs in multimodal tasks that integrate visual and textual understanding, such as visual question answering (Song et al., 2022), image captioning (Dai et al., 2023a), embodied agents (Ma et al., 2025) and document understanding (Luo et al., 2024). Despite their growing deployment, gaps in understanding VLMs' limitations highlight the need for comprehensive evaluation.

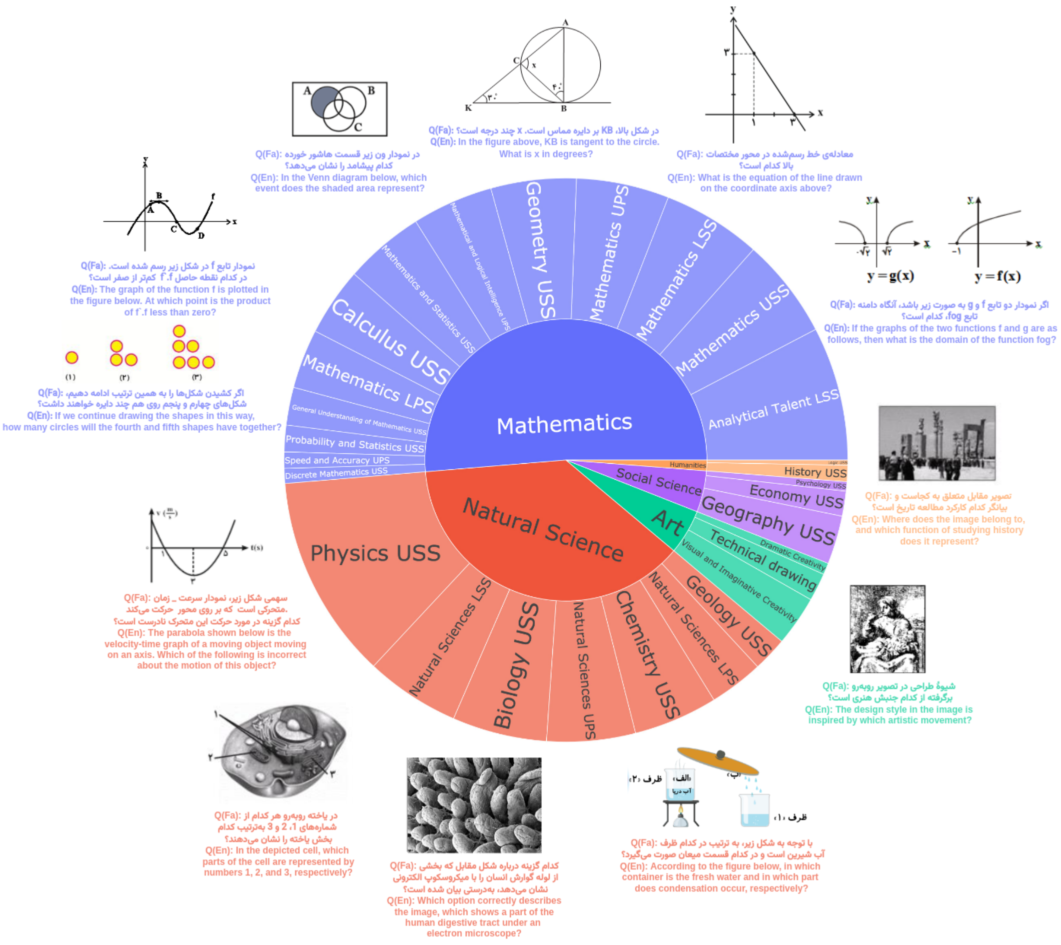
Several benchmarks have been developed to assess VLMs, each addressing different evaluation aspects. MMMU (Yue et al., 2024), derived from college exams, quizzes, and textbooks, is designed to evaluate models on English-language exam questions. BLINK (Fu et al., 2024) focuses on assessing models' performance on tasks that are intuitive for humans, such as

[*] These authors contributed equally to this work and are considered joint second authors. The order is listed randomly to reflect their equal contributions.

[†] These authors contributed equally to this work and are considered joint third authors. The order is listed randomly to reflect their equal contributions.

[‡] These authors contributed equally to this work and are considered joint corresponding authors. The order of corresponding authors is listed randomly to reflect their equal contributions.

visual similarity. MathVista (Lu et al., 2024) specializes in mathematical problem-solving and visual tasks, including tables and bar charts. AI2D (Kembhavi et al., 2016) facilitates question-answering based on diagrams, while MEGA-Bench (Chen et al., 2024b) covers a diverse set of tasks, spanning coding, games, and scientific inquiries. Despite progress in VLM evaluation, existing benchmarks remain predominantly English-centric. Moreover, linguistic and cultural differences underscore the necessity of benchmarks that are natively developed for each language rather than adapted through translation. This creates a pressing need for VLM benchmarks in languages beyond English, including Persian.



Q(Fa): در نمودار ون زیر قسمت هاشور خورده کدام پیشامد را نشان می‌دهد؟
Q(En): In the Venn diagram below, which event does the shaded area represent?

Q(Fa): در شکل بالا، KB بر دایره مماس است. x چند درجه است؟
Q(En): In the figure above, KB is tangent to the circle. What is x in degrees?

Q(Fa): معادله‌ی خط رسم‌شده در محور مختصات بالا کدام است؟
Q(En): What is the equation of the line drawn on the coordinate axis above?

Q(Fa): نمودار تابع f در شکل زیر رسم شده است. در کدام نقطه حاصل f`.f کمتر از صفر است؟
Q(En): The graph of the function f is plotted in the figure below. At which point is the product of f`.f less than zero?

Q(Fa): اگر نمودار دو تابع f و g به صورت زیر باشد، آنگاه دامنه تابع fog، کدام است؟
Q(En): If the graphs of the two functions f and g are as follows, then what is the domain of the function fog?

Q(Fa): اگر کشیدن شکل‌ها را به همین ترتیب ادامه دهیم، شکل‌های چهارم و پنجم روی هم چند دایره خواهند داشت؟
Q(En): If we continue drawing the shapes in this way, how many circles will the fourth and fifth shapes have together?

Q(Fa): تصویر مقابل متعلق به کجاست و بیانگر کدام کارکرد مطالعه تاریخ است؟
Q(En): Where does the image belong to, and which function of studying history does it represent?

Q(Fa): سهمی شکل زیر، نمودار سرعت _ زمان متحرکی است که بر روی محور حرکت می‌کند. کدام گزینه در مورد حرکت این متحرک نادرست است؟
Q(En): The parabola shown below is the velocity-time graph of a moving object moving on an axis. Which of the following is incorrect about the motion of this object?

Q(Fa): شیوهٔ طراحی در تصویر روبه‌رو برگرفته از کدام جنبش هنری است؟
Q(En): The design style in the image is inspired by which artistic movement?

Q(Fa): در یاخته روبه‌رو هر کدام از شماره‌های 1، 2 و 3 به‌ترتیب کدام بخش یاخته را نشان می‌دهند؟
Q(En): In the depicted cell, which parts of the cell are represented by numbers 1, 2, and 3, respectively?

Q(Fa): کدام گزینه درباره شکل مقابل که بخشی از لوله گوارش انسان را با میکروسکوپ الکترونی نشان می‌دهد، به‌درستی بیان شده است؟
Q(En): Which option correctly describes the image, which shows a part of the human digestive tract under an electron microscope?

Q(Fa): با توجه به شکل زیر، به ترتیب در کدام ظرف آب شیرین است و در کدام قسمت میعان صورت می‌گیرد؟
Q(En): According to the figure below, in which container is the fresh water and in which part does condensation occur, respectively?

Figure 1: Overview of dataset and some sample questions from different tasks.

Persian benchmarks have largely concentrated on image captioning and visual question answering. However, they lack coverage of more complex human-level reasoning, such as mathematical problem-solving, spatial reasoning, and academic or scientific exam questions. Furthermore, most existing benchmarks focus primarily on text-based tasks and LLMs, rather than offering a comprehensive evaluation of VLMs.

Existing benchmarks for VLMs often prioritize limited aspects, such as perception or reasoning, mainly for high-resource languages like English, overlooking the challenges posed by diverse linguistic and cultural contexts. In the case of Persian, a relatively low-resource language, these challenges are more pronounced, with available benchmarks failing to account for its unique linguistic features. Farsi et al. (2025) has several limitations, including its reliance on translations from English datasets, automated question generation, and a lack of complex, domain-specific reasoning tasks. Moreover, it does not evaluate reasoning in scientific disciplines such as mathematics or physics, which require structured logical thought. The use of translated datasets introduces cultural misalignment, which undermines the accuracy of evaluations. These limitations highlight the need for more comprehensive evaluation frameworks.

To address these gaps, we introduce **MEENA** (**M**ultimodal **E**ducational **E**xams for **N**-level **A**ssessment), the first dataset designed to evaluate Persian VLMs across scientific, reasoning, and human-level understanding tasks. The name MEENA (Mina) was chosen for this dataset due to its significance in Persian, where ”Mina” refers to glass, and Mina-kari is a traditional art form. This aligns with the dataset's multimodal nature, which includes a subset dedicated to art-related questions. Our dataset comprises approximately 7,500 Persian and 3,000 English questions, covering a wide range of topics such as reasoning, mathematics, physics, diagrams, charts, and Persian art and literature. This benchmark spans diverse subject areas across educational levels, from primary to upper secondary school, providing a comprehensive framework to assess the capabilities of VLMs. Key contributions of our work include:

(1) The first comprehensive Persian multimodal dataset for scientific and art exams, addressing the limitations of previous benchmarks by incorporating scientific and domain-specific tasks.

(2) Extensive experimentation covering a wide range of model evaluation scenarios, including Zero-Shot, Few-Shot, First Describe and focusing on the impact of visual input in different contexts (e.g., ”Wronge Image” and ”Without Image”).

(3) A diverse set of questions varying in difficulty, topic, and format to test model across multiple contexts, such as reasoning, mathematics, physics, diagrams, charts, and Persian art and literature.

(4) Rich metadata, including difficulty levels, descriptive answers, and human performance enabling detailed performance analysis across various dimensions.

(5) Original Persian data that preserves cultural nuances, ensuring accurate evaluation in a culturally relevant context.

(6) A bilingual framework for cross-linguistic model evaluation, enhancing the assessment of performance in both Persian and English, and providing insights into cross-linguistic generalization.

Our dataset and code are available on HuggingFace and GitHub, respectively. Additionally, we set up a leaderboard on HuggingFace to stay updated with the performance of other models.

2 Related Works

2.1 Vision Language Models

Multimodal vision-language models (VLMs) have emerged at the intersection of computer vision and natural language processing, allowing machines to interpret both visual and textual modalities (Li et al., 2025a). The limitations of large language models (LLMs) in handling single-modality data, particularly in capturing real-world information that requires multi-modal perception, have driven researchers to develop VLMs (Li et al., 2025b; Xu et al., 2024). This has led to the rise of various models, including closed-source options like GPT-4o (Hurst et al., 2024), Gemini (Team et al., 2023), and Claude (cla), as well as open-source models such as DeepSeek-VL2 (Wu et al., 2024), InstructBLIP (Dai et al., 2023a), and Qwen2.5-VL (Bai et al., 2025). VLMs are increasingly applied in generative AI systems (Abootorabi et al., 2025a), retrieval-augmented generation (RAG) systems (Abootorabi et al., 2025b), education, and healthcare (Hartsock & Rasool, 2024).

2.2 VLM Evaluation Benchmarks

Despite significant advancements, current VLMs still struggle with certain categories of visual tasks, such as visual arithmetic (Huang et al., 2025) (including geometric problem-

Dataset	Languages	VU Tasks	Type & # Sample (Img)	Access	Metadata			
					Desc. Ans.	Diff. Lev.	Trap	# Tasks (Subtasks)
MMT-Bench	Eng.	MCQA	gen: 31.3K	Open	×	×	×	32 (162)
MMMU	Eng.	MCQA, AM	orig: 11.5K	Open	18%	✓	×	30 (183)
Farsi et. al.	Per., Eng.	MCQA, AM	tran: 7.7K(1K), gen: 70K(7K), orig: 0.6K+?	Closed	×	×	×	5
ParsVQA-Caps	Per.	AM, IC	orig: 27.5K(18.5K)	Open	×	×	-	11
PICD*	Per.	IC	orig: 41K(1.5K)	Open	-	-	-	5
CFF**	Per.	OCR	tran: 124K	Open	-	-	-	1
Persian-OCR	Per.	OCR	orig: 33K	Open (7K)	-	-	-	2
MEENA (Ours)	Per., Eng.	MCQA	orig(Per.): 7.4K (tran(Eng.): 3K)	Open	✓	✓	✓	27

Table 1: A summary of various English and Persian VLM benchmarks, detailing the supported languages, vision understanding tasks, type and number of questions, accessibility, metadata (descriptive answers, difficulty levels, traps), and number of tasks (subtasks). As trapped questions are not defined for questions other than multiple choice, we marked those fields with a hyphen (-). The same is applied for descriptive answers of questions other than multiple choice and answer matching. VU: Vision Understanding, Desc. Ans.: Descripive Answer, Diff. Lev.: Difficulty Level, # Img: Number of images, Eng: English, MCQA: Multiple Choice Question Answering, gen: Generated Questions, orig: Original Questions, tran: Translated Questions, IC: Image Captioning, OCR: Optical Character Recognition, AM: Answer Matching. In Answer Matching, the answer of question could be short-form, long-form, number, or yes/no.
*Persian Image Captioning Dataset (Malekzadeh Lashkaryani, 2021)
**COCO-Flickr Farsi (Kanaani & Ayoubi, 2021)

solving (Gao et al., 2023)) and spatial reasoning, which encompasses spatial relations, orientation, and navigation (Stogiannidis et al., 2025; Chen et al., 2024a). To evaluate VLM performance in these areas, various benchmarks have been introduced. MMMU (Yue et al., 2024) provides a benchmark featuring multiple-choice and open-ended questions designed to assess VLMs' perception, knowledge, and reasoning abilities. MMT-Bench (Ying et al., 2024) evaluates VLMs across 32 tasks requiring expert knowledge, visual reasoning, and localization. Additionally, Stogiannidis et al. (2025); Chen et al. (2024a) present benchmarks specifically focused on assessing VLMs' spatial reasoning capabilities.

Despite existing benchmarks for VLM evaluation, few are designed to assess performance in low-resource languages such as Persian, particularly in scientific knowledge and visual reasoning. COCO-Flickr Farsi (Kanaani & Ayoubi, 2021) and Persian OCR dataset (Aasdi, 2020) address image captioning and optical character recognition. ParsVQA-Caps (Mobasher et al., 2022) provides a visual question answering task in which questions are generated using templates and by human annotators from images gathered from the web.

Ghahroodi et al. (2024) introduces a large-scale, culturally grounded benchmark with 20,805 questions across 38 tasks, enabling rigorous and contamination-free evaluation of LLMs in Persian. Although it covers scientific and reasoning aspects, it lacks visual components and corresponding analysis. Farsi et al. (2025) provides a valuable benchmark with five distinct sets of questions, including visual abstract reasoning, word-image puzzles to assess models' ability to combine visual information with linguistic interpretation, and Iran-places to measures the models' knowledge of notable places in Iran. Despite these valuable questions, the benchmark lacks coverage of most aspects of scientific reasoning (Ma et al., 2024). While it features abstract reasoning questions, it does not evaluate reasoning in many scientific domains such as mathematics or physics, which requires structured logical thinking. This limitation reduces its effectiveness in assessing models' ability to solve complex scientific problems. Furthermore, the dataset lacks diversity in its task range, limiting the subjects on which models can be tested. Using non-original questions translated from English, and questions generated by LLMs to assess VLMs could lead to inaccurate evaluation of models in Persian language (Al Faraby et al., 2024). Additionally, since the dataset is closed-source, its applicability in evaluating other VLMs remains limited. A comparison of a number of Persian and English datasets is provided in Table 1.

3 Dataset

The MEENA benchmark offers a robust collection of data designed to evaluate vision-language models (VLMs) with Persian language support, focusing on multiple-choice question answering. This dataset spans a wide array of disciplines and educational levels, assessing a range of cognitive skills, including reasoning, knowledge application, and comprehension. It is derived from Iran's 12-year educational framework, which consists of 6 years of primary education—divided into lower primary (LP, years 1–3) and upper primary (UP, years 4–6)—and 6 years of secondary education, split into lower secondary (LS, years 7–9) and upper secondary (US, years 10–12).

3.1 Data Compilation

The dataset primarily originates from two sources: (1) the "Pellekan Yadgiri" (Learning Ladder) platform, operated by the Kanoon Farhangi Amoozesh (Cultural Educational Institute) in Iran, which provides educational resources and standardized exercises, and (2) a curated selection of questions from online sources, including items from the Iranian national university entrance exams. The compilation process involved several steps: **(1) Extraction and Cleaning**: Parsed HTML data to extract question attributes, removed questions with tables or explanatory answers, and deduplicated entries. **(2) Image Processing**: Retained only questions with visual elements, categorized as: (i) questions with a single image, (ii) choices with a single image, or (iii) both question and choices with images. For cases with multiple images, these were merged into a single image to ensure compatibility across VLMs, as some models cannot process multiple inputs. Examples are provided in the appendix. **(3) Content Filtering**: Excluded categories with insufficient visual questions (e.g., literature). **(4) Diversity and Contribution**: Questions stem from a broad pool of educators, reducing individual bias and enhancing variety. The dataset is licensed under Creative Commons No Derivatives (CC ND). Sampling was weighted using the formula $1/\text{weight}^{1/4}$, where weight denotes the number of questions per category, tuned to $1/4$ to address data imbalance while preserving diversity. Uniform sampling or fixed-size subsets were avoided to maintain fairness across categories with varying question counts. A bilingual subset of 3,067 questions (547 from online sources and 2,520 from Pellekan Yadgiri) was created by translating items with Persian text in images into English, retaining only those with pure English or non-text visuals. Detailed examples of the MEENA dataset are included in Appendix A.

3.2 Metadata Details

The Pellekan Yadgiri subset, forming the bulk of the MEENA benchmark, includes rich metadata to support in-depth analysis:

- **Educational Level**: Tags questions to LP, UP, LS, or US, aligning difficulty with expected knowledge at each stage.
- **Difficulty Rating**: Assigns one of five levels—easy, moderately easy, medium, moderately hard, hard—for granular performance assessment.
- **Answer Explanations**: Provides detailed reasoning for each correct answer, aiding comprehension and evaluation.
- **Trap Indicators**: Flags questions with misleading "trap" choices, often rated as harder, to study reasoning pitfalls.
- **Student Success Rate**: Records the percentage of students answering correctly, offering a human performance baseline.
- **Subject Breakdown**: Organizes questions into precise topics (e.g., "Mathematics → Algebra → Equations"), enabling targeted analysis.
- **Creation Year**: Tracks the year of question design, revealing trends in complexity over time.

This metadata enables comparisons between human and VLM performance, highlighting strengths and weaknesses in reasoning, trap avoidance, and topic-specific proficiency.

3.3 Statistical Overview

The dataset comprises 7,483 multiple-choice questions: 6,936 from Pellekan Yadgiri and 547 from online sources. It covers domains such as humanities, mathematics, sciences, and reasoning skills, with 6,936 questions linked to human performance data (Pellekan Yadgiri only) and a subset featuring trap elements.

3.4 Translation Process

To extend our primarily Persian dataset into an English counterpart, we adopted a systematic translation pipeline that combines both automated methods and quality checks. Our main translation engine is **GPT-4o**, configured to handle multi-sentence and domain-specific text.

Evaluation Methodology: We applied an **LLM-as-a-Judge** approach, inspired by recent studies Feng et al. (2024); Gu et al. (2025); Zhu et al. (2025); Zheng et al. (2023), in which a large language model (GPT-4o in an evaluator mode) directly compares the translated text to the original Persian input. This model provides a semantic alignment score on a scale from 1 to 5, thus going beyond token matching to incorporate context-aware judgments about meaning preservation and fluency.

Selection Criterion: All translated samples scoring **4 or higher** on the 1–5 scale were retained for the final English dataset. Samples below this threshold underwent additional review or revision to address discrepancies. This filtering ensures that only high-quality English renditions of Persian questions persist, resulting in a consistent, reliable dataset suitable for cross-lingual vision-language model evaluations.

3.5 Distinguishing Features

The MEENA benchmark excels due to:

- **Broad Scope**: Encompasses diverse fields from analytical reasoning to scientific inquiry across educational stages.
- **Enhanced Metadata**: Offers contextual depth for sophisticated model evaluation.
- **Persian Authenticity**: Retains original Persian content with cultural relevance, avoiding translation artifacts.

4 Experiments

4.1 Experiment Overview

We analyze two languages (Persian and English) and classify each question into three different cases based on the presence of images. Specifically:

- **Questions with images:** Only the question prompt contains images.
- **Choices with an image:** Only the answer options (choices) contain images.
- **Both inquiries and selections involving pictures:** Images are present in both the question and its multiple-choice options.

We evaluate **GPT-4o** and **GPT-4o-mini** (OpenAI, 2024), **GPT-4-Turbo** (OpenAI, 2023), **Gemini-2.0-flash** (DeepMind, 2023), and **InstructBLIP-T5** (Dai et al., 2023b) on **Persian** and **English** data.

To determine how visual information affects the model's performance, each of these three cases is examined independently. We further design five experimental settings (Zero-Shot, In-Context Learning, First Describe, Wrong Image, Without Image) to isolate the impact of different multimodal cues and prompting strategies. for further details about the models and the rationality behind the experiment settings, see Appendix B.

4.2 Experimental Design

Below, we formalize each of our five main experiment types using a uniform notation. Let q_* be the textual question (in Persian or English), x_* be the (true or substituted) image relevant to q_*, c_* be the correct answer or label we aim to predict, $M(\cdot)$ denote the model's output given specified inputs. For every experiment, the same set of questions is used.

Zero-Shot (ZS). A minimal-guidance setup in which the model receives only the single question-image pair $\left(q_*, x_*\right)$ with no supplemental examples:

$$\hat{c}_* \;=\; M\bigl(q_*,\, x_*\bigr).$$

Here, $\hat{c}_*$ represents the model's direct output under default settings. Concretely, each input prompt includes the text of q_* and the raw image as two distinct inputs. No additional context (such as sample Q&A pairs) is provided. Each pair is processed independently, ensuring no cross-contamination of information between different items.

In-Context Learning (ICL). We provide k example triplets $\{(q_i, c_i)\}_{i=1}^{k}$ as demonstrations immediately before the target query (q_*, x_*):

$$\hat{c}_* \;=\; M\Bigl(\bigl\{(q_i, c_i)\bigr\}_{i=1}^{k},\; q_*, x_*\Bigr).$$

The value of k set to four and is kept consistent within each run. Additionally, the examples were chosen manually, ensuring the examples are informative and relevant to the questions topic.

First Describe (FD). We draw inspiration from chain-of-thought prompting approaches (Wei et al., 2023) that encourage models to generate intermediate reasoning steps in text form before producing a final output. Similar works on multimodal reasoning (Rose et al., 2024; Zhang et al., 2024; Zheng et al., 2024) also motivate explicit step-by-step analysis of visual content. In our adaptation, we create a form of ”visual chain of thought” for each image, aiming to prevent the model from taking shortcuts (i.e., guessing an answer without fully accounting for the image).

Experiments with Mismatched or Missing Images. Before introducing the *Wrong Image* and *Without Image* settings, we note that prior research on multimodal grounding and visual-text alignment has explored techniques such as image substitution or omission to diagnose model dependencies (Hemmat et al., 2024; Favero et al., 2024; Villa et al., 2025; Gunjal et al., 2024; Wang et al., 2024). In our design, we follow similar practices to investigate whether the absence or irrelevance of the image affects a model's predictive outcome. We adopt two settings that vary the presence or correctness of the accompanying image:

Wrong Image (WI). We replace the correct image x_* with an intentionally mismatched or irrelevant image $\hat{x}$ that does not correspond to q_*:

$$\hat{c}_* \;=\; M\bigl(q_*,\, \hat{x}\bigr).$$

All other prompt elements remain unchanged. Each wrong image $\hat{x}$ is drawn from a pool of images that are confirmed to be unrelated to the content of q_*. This ensures the mismatch is unambiguous. The input format (text+image) is kept identical to Zero-Shot, except we swap out the image.

Without Image (WO). We remove the image entirely:

$$\hat{c}_* \;=\; M\bigl(q_*,\, x_* = \varnothing\bigr).$$

In practice, the model is given only the text of q_*, and references to an image are either omitted or replaced with a placeholder (e.g., “[No Image Provided]”) depending on how the prompts are typically structured. The rest of the setup, including question style and domain, remains identical to Zero-Shot.

Methods&Datasets	**Zero Shot**	**ICL**	**First Describe**	**Wrong Image**	**Without Image**
MEENA Persian Dataset					
GPT-4o-mini	0.310	0.224	0.312	0.221	0.235
GPT-4o	0.413	**0.385**	0.422	**0.247**	**0.292**
GPT-4-Turbo	0.313	0.310	0.295	-	0.213
Gemini-2.0-flash	**0.435**	0.377	**0.504**	0.121	0.267
MEENA English Dataset					
GPT-4o-mini	0.368	0.312	0.361	0.275	0.279
GPT-4o	0.474	0.397	**0.464**	0.269	**0.401**
GPT-4-Turbo	0.440	0.384	0.381	**0.306**	0.304
Gemini-2.0-flash	**0.494**	**0.464**	0.459	0.178	0.311
instructblip-t5	0.226	*	0.193	0.197	*
Art Persian Dataset					
GPT-4o-mini	0.323	0.250	0.248	**0.193**	0.206
GPT-4o	**0.354**	0.239	0.374	0.171	0.182
GPT-4-Turbo	0.305	0.305	0.265	-	0.186
Gemini-2.0-flash	0.297	**0.318**	**0.387**	0.122	**0.244**
Art English Dataset					
GPT-4o-mini	0.343	0.276	0.301	0.241	0.217
GPT-4o	0.372	0.311	0.406	0.230	0.232
GPT-4-Turbo	0.336	0.374	**0.334**	0.197	**0.294**
Gemini-2.0-flash	**0.376**	**0.372**	0.329	0.151	0.159
instructblip-t5	0.274	*	0.266	**0.274**	*

Table 2: Accuracy comparison of different models across various tasks (Zero Shot, In-Context Learning, First Describe, Wrong Image, and Without Image) on the MEENA and Art datasets in both Persian and English. An asterisk (*) in the table indicates that the model does not support the corresponding setting.

4.3 Answer Extraction

To assess model performance, it is essential to identify the option selected by the model in its generated response and use it to compute accuracy. To achieve this, we implement a two-stage framework. In the first stage, we apply regex-based pattern matching to extract explicit statements, such as ”The correct answer is option 2.” When these predefined rule-based patterns match, we can confidently extract the model’s selected option. However, in approximately half of the cases, regex patterns do not yield a match. Furthermore, depending on the nature of the experiment such as scenarios where no image is provided, the model may correctly infer the absence of an image and generate a response like ”An image is required to answer this question.” To handle such cases, we leverage LLM as a judge, utilizing the GPT-4o-mini model to infer the selected option when explicit patterns are absent. This model also determines whether the response indicates a missing image, assesses instances where the model fails to comprehend the question, and identifies responses that indicate an incorrect image reference.

All prompts used in the experiments, translations, and LLM as a judge are provided in the appendix D.

5 Results and discussions

Knowledge-Based Tasks Consistently Outperform Reasoning Ones: The evaluation presented in Figure 2(a) highlights a performance gap between knowledge-based and reasoning tasks across various models. Knowledge-based tasks consistently outperform reasoning tasks by a significant margin of +10–19% in absolute accuracy. This trend is observed for both English and Persian tasks, though Persian tasks generally exhibit lower accuracy, likely due to differences in training data distribution. These results suggest that while current vision-language models excel at factual recall, they face greater challenges with complex reasoning tasks. Moreover, the performance gap is more pronounced in Persian, indicating that reasoning tasks in this language are more difficult than in English.

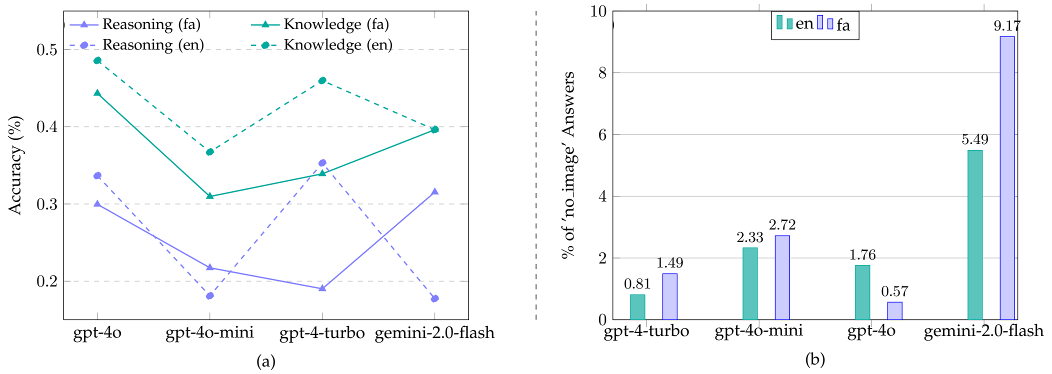


Figure 2: (a) Accuracy comparison of reasoning and knowledge-based tasks across models in English (en) and Persian (fa). Solid lines represent Persian tasks, while dashed lines indicate English tasks. (b) Comparison of 'no image' error rates for English (en) and Persian (fa) inputs. GPT-4-Turbo and GPT-4o maintain consistently low error rates in both languages, while Gemini 2.0 Flash exhibits significantly higher errors, particularly for Persian inputs.

Hallucination Detection Performance with Incorrect Images: Figure 3 compares hallucination detection rates across three vision-language models—Gemini 2.0 Flash, GPT-4, and GPT-4 Mini—on the Art and MEENA datasets in both English and Persian. To evaluate hallucination detection, we replace each query's image with an incorrect one (Section 4.2) and consider a detection successful only if the model identifies the mismatch. Gemini 2.0 Flash consistently achieves higher detection rates than both GPT-4 and GPT-4 Mini across datasets, with a particularly significant performance gap in Persian. The detection rate difference between Gemini 2.0 Flash and GPT-4 Mini on the MEENA dataset is over 400 detections, suggesting that Gemini 2.0 Flash is more robust at recognizing inconsistencies, especially in Persian contexts.

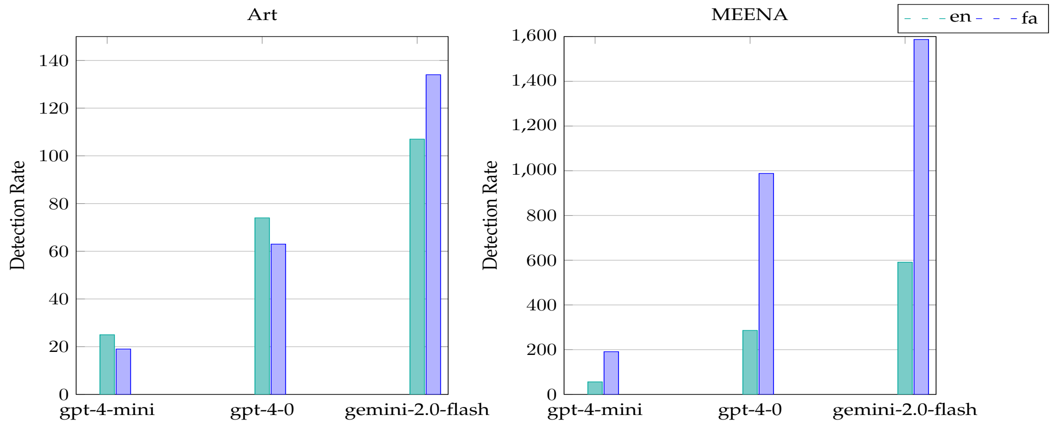


Figure 3: Hallucination detection rates across three vision-language models (GPT-4 Mini, GPT-4, and Gemini 2.0 Flash) on the Art and MEENA datasets, for both English and Persian. The bars represent detection rates for each model in both languages, with a clear performance gap observed in Persian, particularly for Gemini 2.0 Flash.

***No Image* Errors in Image Detection Across Models**: Figure 2(b) illustrates the frequency at which different models mistakenly report the absence of an image, despite one being provided. The chart displays the percentage of 'no image' responses for four models, evaluated on both English and Persian inputs. GPT-4-Turbo and GPT-4o demonstrate relatively low 'no image' error rates across both languages, with English inputs yielding slightly fewer errors than Persian. In contrast, Gemini 2.0 Flash exhibits a markedly higher incidence of 'no image' responses, particularly for Persian inputs, where the error rate reaches 9.17%.

Models Struggle with Higher-Level Questions: Figures 4 and 5 show that as question difficulty increases in the Chemistry and Mathematics tasks of the zero-shot experiment in English, model performance generally declines. While models like GPT-4o-mini and GPT-4-Turbo experience significant drops in accuracy at higher levels, Gemini-2.0-flash maintains relatively consistent performance, particularly in the Mathematics task. In contrast, instructblip-t5 struggles across all levels, especially in the Chemistry task. Further results are provided in the appendix C.

Model	lvl 12	lvl 11	lvl 10	lvl 9	lvl 8	lvl 7	lvl 6	lvl 5	lvl 4	lvl 3	lvl 2
GPT-4o-mini	0.29	0.27	0.32	0.5	0.49	0.57	0.33	0.65	0.57	0.57	0.71
GPT-4o	0.38	0.37	0.48	0.64	0.61	0.68	0.63	0.75	0.77	0.72	0.87
GPT-4-Turbo	0.23	0.32	0.48	0.57	0.56	0.68	0.52	0.68	0.52	0.6	0.81
Gemini-2.0-flash	0.47	0.4	0.72	0.59	0.72	0.75	0.59	0.68	0.66	0.62	0.76
instructblip-t5	0.42	0.16	0.2	0.36	0.28	0.27	0.14	0.42	0.34	0.37	0.42

Figure 4: Heatmap of model accuracy across different levels of the **MEENA English** dataset for the **Chemistry Course/Experimental Science** in the **Zero-shot** experiment.

Model	lvl 12	lvl 11	lvl 10	lvl 9	lvl 8	lvl 7	lvl 6	lvl 5	lvl 4	lvl 3	lvl 2
GPT-4o-mini	0.45	0.23	0.41	0.44	0.29	0.35	0.46	0.33	0.31	0.42	0.36
GPT-4o	0.35	0.41	0.5	0.48	0.37	0.64	0.37	0.55	0.65	0.5	0.6
GPT-4-Turbo	0.25	0.32	0.38	0.44	0.34	0.47	0.48	0.5	0.42	0.5	0.49
Gemini-2.0-flash	0.48	0.57	0.53	0.63	0.53	0.64	0.43	0.5	0.47	0.58	0.57
instructblip-t5	0.16	0.27	0.28	0.14	0.16	0.2	0.2	0.22	0.19	0.14	0.21

Figure 5: Heatmap of model accuracy across different levels of the **MEENA English** dataset for the **Mathematics** in the **Zero-Shot** experiment.

6 Conclusions

In this study, we present MEENA, the first benchmark designed to assess scientific reasoning, problem-solving, and human-level Persian language understanding in VLMs. MEENA comprises multiple-choice questions available in both Persian and English, enriched with extensive metadata, including difficulty levels and descriptive answers. Furthermore, we conducted a series of experiments to analyze different model capabilities, including Zero-Shot, In-Context Learning, First Describe, Wrong Image, and Without Image scenarios. Our evaluation highlights key performance trends across vision-language models. (1) Knowledge-based tasks consistently outperform reasoning-based ones, with a more pronounced gap in Persian. (2) Gemini 2.0-flash surpasses GPT-4 and GPT-4o-Mini in detecting image mismatches, demonstrating greater reliability in mitigating hallucinations, particularly in Persian. (3) GPT-4-Turbo and GPT-4o excel in image presence detection, while Gemini 2.0-flash shows higher 'no image' error rates. (4) Models struggle with higher-level Chemistry and Mathematics questions, with performance declining as complexity increases. These findings emphasize the challenges of complex reasoning and domain-specific knowledge retrieval in both Persian and English for vision-language models.

References

Claude 3.7 sonnet and claude code anthropic. URL https://www.anthropic.com/news/claude-3-7-sonnet.

Amirabbas Aasdi. Persian ocr dateset, 2020. URL https://github.com/persiandataset/Arshasb.

Mohammad Mahdi Abootorabi, Omid Ghahroodi, Pardis Sadat Zahraei, Hossein Behzadasl, Alireza Mirrokni, Mobina Salimipanah, Arash Rasouli, Bahar Behzadipour, Sara Azarnoush, Benyamin Maleki, Erfan Sadraiye, Kiarash Kiani Feriz, Mahdi Teymouri Nahad, Ali Moghadasi, Abolfazl Eshagh Abianeh, Nizi Nazar, Hamid R. Rabiee, Mahdieh Soleymani Baghshah, Meisam Ahmadi, and Ehsaneddin Asgari. Generative ai for character animation: A comprehensive survey of techniques, applications, and future directions, 2025a. URL https://arxiv.org/abs/2504.19056.

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16776–16809, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.861. URL https://aclanthology.org/2025.findings-acl.861/.

Said Al Faraby, Ade Romadhony, et al. Analysis of llms for educational question classification and generation. *Computers and Education: Artificial Intelligence*, 7:100298, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024a.

Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Wang Zhu, Ziyan Jiang, Bohan Lyu, Dongfu Jiang, Xuan He, Yuan Liu, Hexiang Hu, Xiang Yue, and Wenhu Chen. Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks, 2024b. URL https://arxiv.org/abs/2410.10563.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023a. URL https://arxiv.org/abs/2305.06500.

Xuehai Dai, Zhe Yang, Junnan Li, Luowei Zhou, Mohammed Almazrooie, Zhe Gan, Zhangyang Wang, Zicheng Liu, and Linjie Li. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023b.

Google DeepMind. Gemini 1 and 2: Multimodal capabilities and performance, 2023. URL https://deepmind.google/technologies/gemini/. Accessed: 2025-03-28.

Farhan Farsi, Shahriar Shariati Motlagh, Shayan Bali, Sadra Sabouri, and Saeedeh Momtazi. Persian in a court: Benchmarking VLMs in Persian multi-modal tasks. In Wei Emma Zhang, Xiang Dai, Desmond Elliot, Byron Fang, Mongyuan Sim, Haojie Zhuang, and Weitong Chen (eds.), *Proceedings of the First Workshop of Evaluation of Multi-Modal Generation*, pp. 52–56, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.evalmg-1.5/.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding, 2024. URL https://arxiv.org/abs/2403.14003.

Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. M-mad: Multidimensional multi-agent debate framework for fine-grained machine translation evaluation. *arXiv preprint arXiv:2412.20127*, 2024. URL https://arxiv.org/abs/2412.20127.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive, 2024. URL https://arxiv.org/abs/2404.12390.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.

Omid Ghahroodi, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib, Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. Khayyam challenge (persianMMLU): Is your LLM truly wise to the persian language? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=yIEyHP7AvH.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models, 2024. URL https://arxiv.org/abs/2308.06394.

Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: a review. *Frontiers in Artificial Intelligence*, Volume 7 - 2024, 2024. ISSN 2624-8212. doi: 10.3389/frai.2024.1430984. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1430984.

Arshia Hemmat, Adam Davies, Tom Lamb, Jianhao Yuan, Philip Torr, Ashkan Khakzar, and Francesco Pinto. Hidden in plain sight: Evaluating abstract shape recognition in vision-language models. *Advances in Neural Information Processing Systems*, 37:88527–88556, 2024.

Kung-Hsiang Huang, Can Qin, Haoyi Qiu, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. Why vision language models struggle with visual arithmetic? towards enhanced chart and geometry understanding. *arXiv preprint arXiv:2502.11492*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Navid Kanaani and Sajjad Ayoubi. Coco-flickr farsi, 2021. URL https://www.kaggle.com/datasets/navidkanaani/coco-flickr-farsi.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016. URL https://arxiv.org/abs/1603.07396.

Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *arXiv preprint arXiv:2501.02189*, 2025a.

Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges, 2025b. URL https://arxiv.org/abs/2501.02189.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15630–15640, 2024.

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, Aixin Sun, Hany Awadalla, et al. Sciagent: Tool-augmented language models for scientific reasoning. *arXiv preprint arXiv:2402.11451*, 2024.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai, 2025. URL https://arxiv.org/abs/2405.14093.

Arman Malekzadeh Lashkaryani. Persian image captioning dataset, 2021. URL https://www.kaggle.com/datasets/malekzadeharman/persian-image-captioning-dataset.

Shaghayegh Mobasher, Ghazal Zamaninejad, Maryam Hashemi, Melika Nobakhtian, and Sauleh Eetemadi. Parsvqa-caps: A benchmark for visual question answering and image captioning in persian. *people*, 101:404, 2022.

OpenAI. Gpt-4 turbo overview, 2023. URL https://openai.com/blog/new-models-and-developer-products-announced-at-devday. Accessed: 2025-03-28.

OpenAI. Gpt-4o technical report, 2024. URL https://openai.com/index/gpt-4o. Accessed: 2025-03-28.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: Bridging logical gaps with multimodal infillings, 2024. URL https://arxiv.org/abs/2305.02317.

Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. Clip models are few-shot learners: Empirical studies on vqa and visual entailment, 2022. URL https://arxiv.org/abs/2203.07190.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A. Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models, 2025. URL https://arxiv.org/abs/2503.19707.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Andrés Villa, Juan León Alcázar, Motasem Alfarra, Vladimir Araujo, Alvaro Soto, and Bernard Ghanem. Eagle: Enhanced visual grounding minimizes hallucinations in instructional multimodal models, 2025. URL https://arxiv.org/abs/2501.02699.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding, 2024. URL https://arxiv.org/abs/2403.18715.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

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

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2024. URL https://arxiv.org/abs/2302.00923.

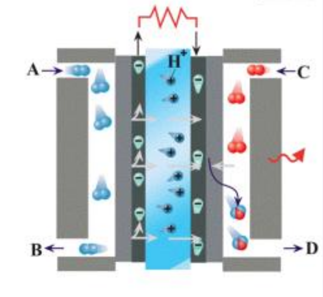
Haojie Zheng, Tianyang Xu, Hanchi Sun, Shu Pu, Ruoxi Chen, and Lichao Sun. Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination, 2024. URL https://arxiv.org/abs/2411.12591.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges, 2025. URL https://arxiv.org/abs/2310.17631.

A Dataset Additional Information

We provide several examples from our dataset, MEENA, showcasing the diversity and structure of the questions included. Each example contains a visual component along with corresponding questions in both Persian and English. These samples illustrate different question formats, such as multiple-choice questions, mathematical problem-solving, and pattern recognition, all integrated with images.



Persian question:

شکل مقابل نشان‌دهنده رایج‌ترین سلول سوختی است. چند مورد از مطالب زیر نادرست هستند؟

آ) در این سلول دو گاز به‌طور کنترل‌شده با یکدیگر وارد واکنش می‌شوند و در حدود ۰۶ درصد از انرژی شیمیایی تولیدی به انرژی الکتریکی تبدیل می‌شود.

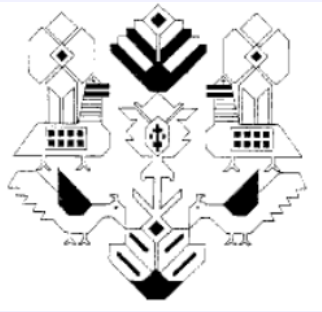
ب) واکنش کلی انجام‌شده در این سلول به‌صورت $۲H_۲(g)+O_۲(g) \rightarrow ۲H_۲O\,(l)$است.

پ) در این سلول جریان الکترون‌ها در مدار بیرونی برخلاف جریان پروتون‌ها در غشای مبادله‌کننده پروتون، از آند به کاتد است.

ت) گاز B همان گاز A است که می‌تواند به عنوان سوخت این سلول به‌طور پیوسته وارد سلول شده و اکسایش یابد.
$(C = 12\,,\, O = 16\;,\; H = 1 : g.mol^{-1})$

۱- ۱
۲- ۲
۳- ۳
۴- ۴

English question:
The figure opposite represents the most common fuel cell. How many of the following statements are incorrect?
A) In this cell, two gases react with each other in a controlled manner, and about 60% of the generated chemical energy is converted into electrical energy.
B) The overall reaction occurring in this cell is represented as $2H_2(g) + O_2(g) \rightarrow 2H_2O\,(l)$.
C) In this cell, the flow of electrons in the external circuit is from anode to cathode, opposite to the flow of protons in the proton exchange membrane.
D) Gas B is the same as gas A, which can continuously enter the cell as fuel and be oxidized.
1) 1 2) 2 3) 3 4) 4



Persian question:

طرح رو به رو با موضوع مرغ و درخت، از دست‌بافت‌های سنتی کدام اقوام است؟

۱- ایل عرب - خوزستان

۲- ایل بهارلو - فارس

۳- لک - کرمانشاه

۴- ترک - همدان

English question:
The design featuring a chicken and a tree, shown in front, is a traditional handmade craft of which ethnic group?

1) Arab Tribe - Khuzestan

2) Baharlu Tribe - Fars

3) Lak - Kermanshah

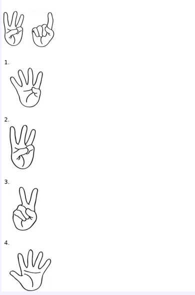
4) Turk - Hamedan

Figure 6: Sample of MEENA questions

Question

مجموع انگشتان باز دست راست و انگشتان باز دست چپ شکل زیر، در کدام گزینه به صورت صحیح آمده است؟

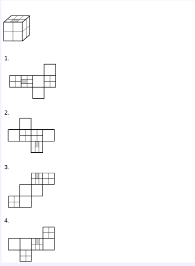
Which option correctly states the sum of the open fingers of the right hand and the open fingers of the left hand in the figure below?



Question:

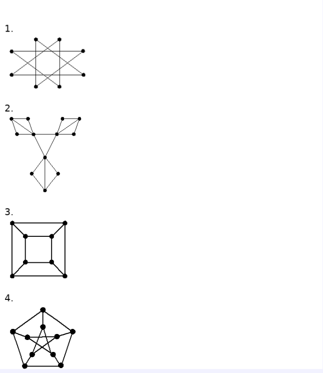
ا ز کدام شکل گسترده مکعبی با نمای روبه‌رو حاصل می‌شود؟ پشت برگه‌ها کاملاً سفید است.

From which unfolded shape is the cube with the front view obtained? The back of the sheets is completely white.



Question

عدد احاطه‌گری کدام گزینه متفاوت از گزینه‌های دیگر است؟



Question:

غذای کدام جانور، میوه یا دانه نیست ؟

Which animal’s food is not fruit or seeds?

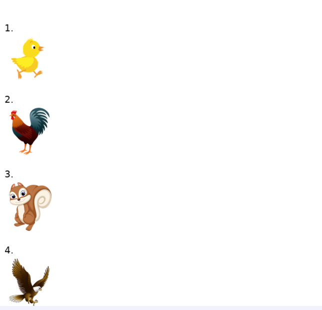
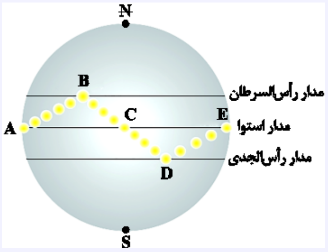


Figure 7: Sample of MEENA questions with picture in choices



Persian question:

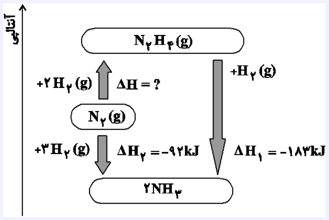
اگر خط g مطابق شکل زیر در نقطه $A(۲,۳)$ بر نمودار$f(x)$ مماس و $f'(۲) = ۲$ باشد، آن‌گاه عرض از مبدأ خط g کدام است؟

-۱ -۱

۲ -۲

۱ -۳

-۳ -۴



Persian question:

با توجه به نمودار زیر برای تولید ۱۶۰ گرم هیدرازین از گازهای نیتروژن و هیدروژن، چند کیلوژول انرژی لازم است؟

$(H = 1\ ,\quad N = 14 : g\,.\,mol^{-1})$

۵۲۲/۵ -۱

۴۵۵ -۲

۹۱۰ -۳

۱۳۷۵ -۴

Figure 8: Sample of MEENA questions including Persian texts in picture

B Experiments Additional Information

2.1 Models Used

We evaluate the following models in our experiments:

- **GPT-4o** and **GPT-4o-mini**: Larger and smaller variants of OpenAI's GPT-4-based architecture capable of processing text, images, and audio, designed for real-time multimodal interaction (OpenAI, 2024).
- **GPT-4-Turbo**: An optimized variant of GPT-4, developed by OpenAI, suited for interactive dialogue with improved cost and performance characteristics (OpenAI, 2023).
- **Gemini-2.0-flash**: A multimodal vision-language model developed by Google DeepMind, trained to process and integrate text, image, and video inputs efficiently (DeepMind, 2023).
- **InstructBLIP-T5**: A T5-based vision-language model that incorporates instruction tuning and visual grounding to handle complex multimodal tasks (Dai et al., 2023b).

By evaluating all models on the same tasks and under each of the five experimental settings, we can measure their relative strengths and weaknesses in multimodal reasoning.

2.2 Experimental Cases and Motivations

We apply each of the five experiment types (ZS, ICL, FD, WI, WO) to the three image-based question categories introduced earlier: (1) questions with images, (2) choices with images, and (3) both questions and choices containing images. The rationale for each experiment type is as follows:

- **Zero-Shot (ZS)**: Establishes a baseline for model performance without contextual examples.
- **In-Context Learning (ICL)**: Investigates whether a few-shot prompt improves multimodal understanding.
- **First Describe (FD)**: Tests whether forcing a detailed image description yields more accurate reasoning.
- **Wrong Image (WI)**: Assesses how reliant the model is on correct image cues (detecting mismatches, etc.).
- **Without Image (WO)**: Shows performance under pure text-only conditions, contrasting it with results that use images.

C Results

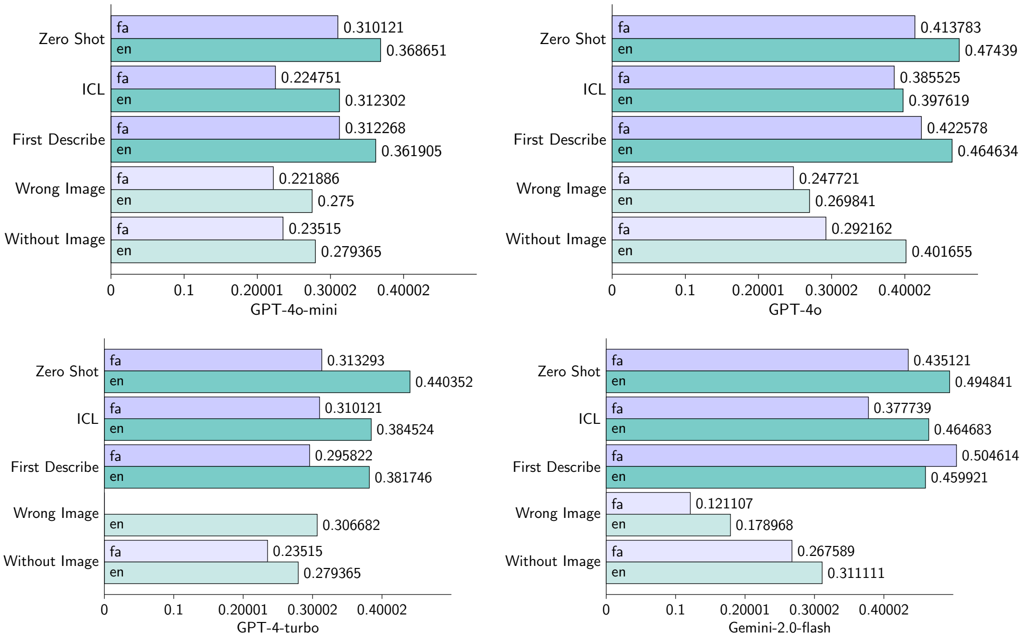


Figure 9: Comparison of Farsi and English performance across different experiments and models on the **MEENA** dataset.

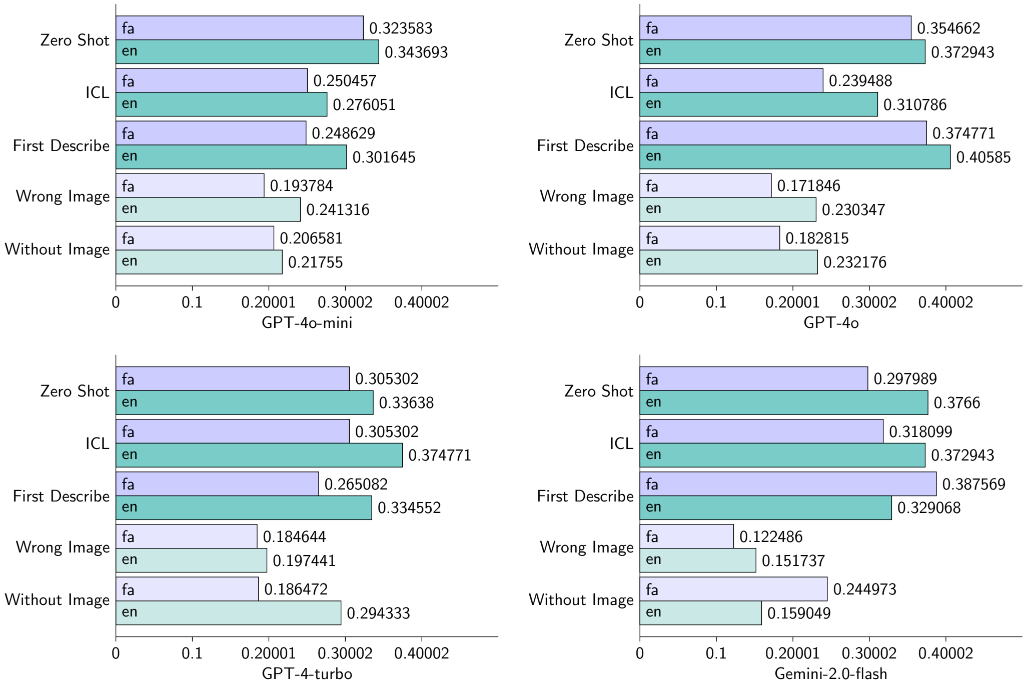


Figure 10: Comparison of Farsi and English performance across different experiments and models on the **Art** dataset.

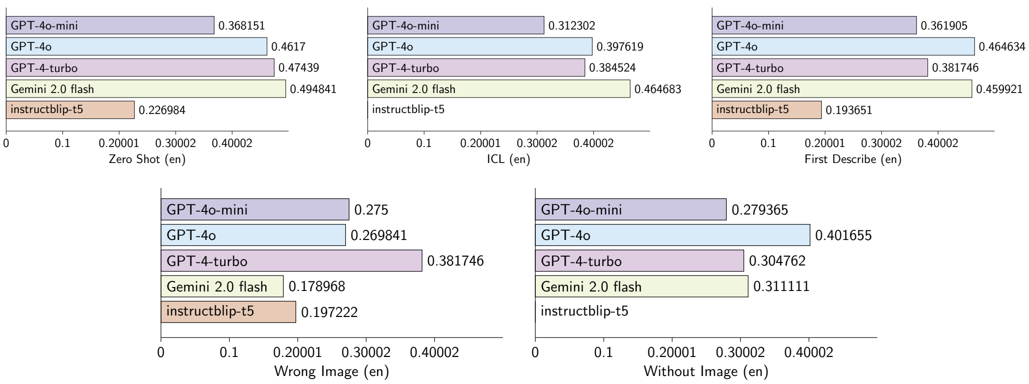


Figure 11: Performance comparison of each model across experiments on the **MEENA English** dataset

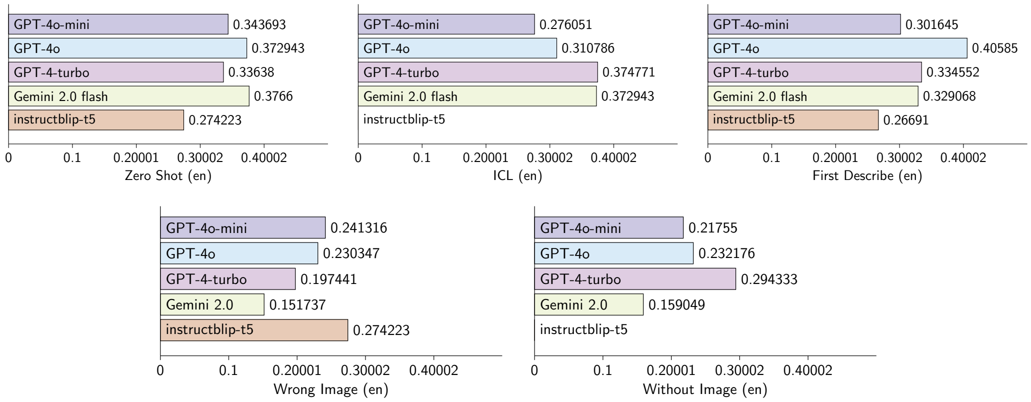


Figure 12: Performance comparison of each model across experiments on the **Art English** dataset

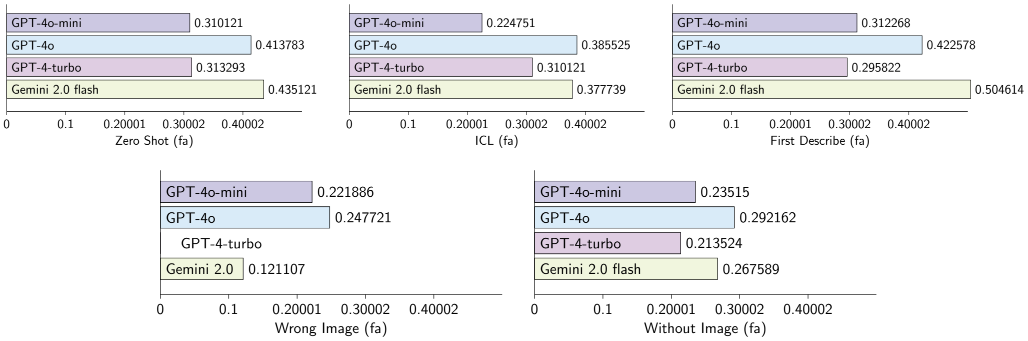


Figure 13: Performance comparison of each model across experiments on the **MEENA Farsi** dataset

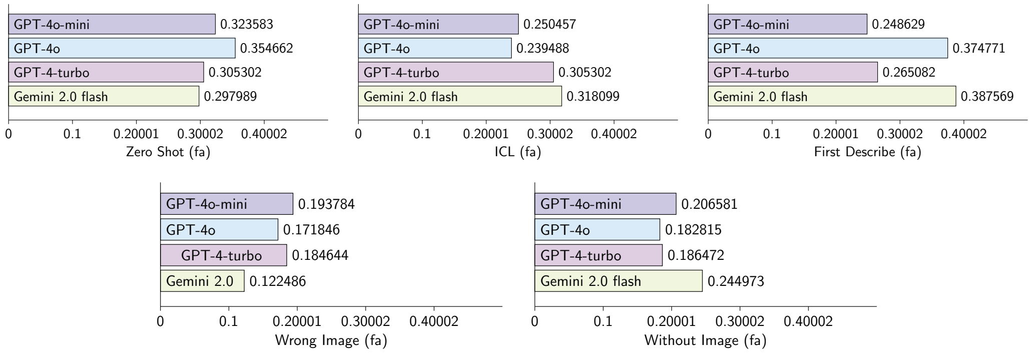


Figure 14: Performance comparison of each model across experiments on the **Art Farsi** dataset

Methods	**Mathematics**	**Natural Science**	**Social Science**	**Humanities**	**Other**
Zero-Shot					
GPT-4o-mini	0.346418	0.619835	0.553571	0.417815	0.199475
GPT-4o	0.460041	**0.675**	**0.625**	0.5397	0.263298
GPT-4-Turbo	0.39417	0.467532	0.363636	0.477707	0
Gemini-2.0-flash	**0.490677**	0.661157	**0.625**	**0.540827**	**0.32021**
instructblip-t5	0.179588	0.330579	0.303571	0.279958	0.178478
First-Descibe					
GPT-4o-mini	0.348381	0.528926	0.446429	0.408271	0.217848
GPT-4o	0.436475	0.675	**0.678571**	**0.535408**	0.263298
GPT-4-Turbo	0.339549	0.545455	0.517857	0.465536	0.215223
Gemini-2.0-flash	**0.478901**	**0.570248**	0.625	0.487805	**0.28084**
instructblip-t5	0.155054	0.330579	0.25	0.26193	0.076115
ICL					
GPT-4o-mini	0.320903	0.454545	0.392857	0.33298	0.181102
GPT-4o	0.421001	0.528926	**0.535714**	0.40403	**0.257218**
GPT-4-Turbo	0.33366	0.528926	0.5	0.474019	0.23622
Gemini-2.0-flash	**0.498528**	**0.586777**	0.482143	**0.50053**	0.244094
Wrong Image					
GPT-4o-mini	0.270854	0.454545	0.464286	0.342524	0.034121
GPT-4o	0.274779	0.404959	0.339286	0.340403	0.028871
GPT-4-Turbo	0.245752	0.25	0.272727	0.364246	0
Gemini-2.0-flash	0.165849	0.289256	0.303571	0.236479	0.018373
instructblip-t5	0.180569	0.264463	0.285714	0.265111	0.03937
Without Image					
GPT-4o-mini	0.276742	0.446281	0.392857	0.358431	0.020997
GPT-4o	0.274779	0.404959	0.339286	0.340403	0.028871
GPT-4-Turbo	0.26791	0.429752	0.446429	0.430541	0.031496
Gemini-2.0-flash	0.31894	0.330579	0.446429	0.415695	0.005249

Table 3: Accuracy comparison of different models across different experiments and course subjects on the **English MEENA** dataset.

Methods	**Mathematics**	**Natural Science**	**Social Science**	**Humanities**	**Other**
Zero-Shot					
GPT-4o-mini	0.284702	0.439437	0.380435	0.341637	0.202091
GPT-4o	0.377134	0.55493	0.554348	0.462989	**0.261324**
GPT-4-Turbo	0.252174	0.422535	0.554348	0.379359	0.214286
Gemini-2.0-flash	**0.414493**	**0.577465**	**0.586957**	**0.476868**	0.229965
First-Describe					
GPT-4o-mini	0.270946	0.401709	0	0.346692	0
GPT-4o	0.385829	0.6	**0.630435**	0.469395	0.249129
GPT-4-Turbo	0.238614	0.45283	0.484848	0.354093	0.203833
Gemini-2.0-flash	**0.509501**	**0.566197**	0.565217	**0.541281**	**0.250871**
ICL					
GPT-4o-mini	0.242512	0.267606	0.271739	0.206762	0.183074
GPT-4o	0.37037	0.490141	0.521739	0.411388	**0.254355**
GPT-4-Turbo	0.249919	0.461972	0.51087	0.372598	0.203833
Gemini-2.0-flash	**0.509501**	**0.566197**	**0.565217**	**0.541281**	0.250871
Wrong Image					
GPT-4o-mini	0.202576	0.35493	0.25	0.260142	0.052265
GPT-4o	0.22754	0.366667	0.378049	0.256077	0.115108
Gemini-2.0-flash	0.094042	0.267606	0.326087	0.148754	0.008711
Without Image					
GPT-4o-mini	0.202254	0.397183	0.402174	0.286833	0.033101
GPT-4o	0.226	0.442623	0	0.346743	0
GPT-4-Turbo	0.150081	0.383099	0.456522	0.291815	0.029617
Gemini-2.0-flash	0.234461	0.388732	0.423913	0.330249	0.04007

Table 4: Accuracy comparison of different models across different experiments and course subjects on the **Farsi MEENA** dataset.

D Prompts

LLM as a Judge(Translation)

Welcome to the Translation Quality Assessment Tool. This tool is designed to evaluate the quality of translations from Persian to English. Please read each translated text carefully and use the rubric below to assign a score based on how well the translation preserves the meaning of the original text.

Scoring Rubric:
- 5 - Excellent: The translation conveys all aspects of the original meaning without any noticeable errors.
- 4 - Good: Minor errors are present but do not alter the fundamental meaning.
- 3 - Acceptable: Some parts of the meaning are lost or altered, but the overall intent is still recognizable.
- 2 - Poor: Significant portions of the meaning are incorrect or missing.
- 1 - Unacceptable: The translation fails to convey the original meaning.

Please focus on the semantic accuracy of the translation. Minor grammatical or syntactic errors should be overlooked unless they significantly impact the overall understanding of the text.

LLM as a Judge(Answer Extraction)

Given a question and its answer, identify the selected option (1, 2, 3, or 4). If the choice is explicitly mentioned by number or implicitly indicated by the content, return the corresponding numerical option. If the answer mentions it does not have the image, return ”no_image.” If the answer mentions it does not understand the image, return ”cannot_understand.” If the answer mentions that the image is wrong or not relevant to the question, return ”wrong_image.” If the selection cannot be determined or is ambiguous, return ”unknown.” Output only the number (1, 2, 3, or 4), ”no_image”, ”cannot_understand”, ”wrong_image”, or ”unknown” without any explanations.

Persian to English translation prompts

Please accurately translate the following multiple-choice question from Persian to English (Both Question and corresponding options). Use precise and professional terminology, especially if specialized terms are involved. The question may be related to fields such as art, chemistry, physics, mathematics, biology, linguistics, geology, or other domains. Ensure that the translation is accurate and maintains the original meaning.
Question:{mcq_question}

Please answer like this template:
Question:
Choice 1:
Choice 2:
Choice 3:
Choice 4:

English First Describe Prompt

Below, you can see multiple-choice questions (with answers).
Question: {mcq question}
Choices:
1)
2)
3)
4)
Answer: Let's first describe the image carefully and provide all its details, then answer the question

English ICL Prompt

Below, you can see multiple-choice questions (with answers).
Question: {mcq question-shot1}
Choices:
1)
2)
3)
4)
Answer: {shot1 answer}
Question: {mcq question-shot2}
Choices:
1)
2)
3)
4)
Answer: {shot2 answer}
Question: {mcq question-shot3}
Choices:
1)
2)
3)
4)
Answer: {shot3 answer}
Question: {mcq question-shot4}
Choices:
1)
2)
3)
4)
Answer: {shot4 answer}
Question: {mcq question}
Choices:
1)
2)
3)
4)
Answer:

Model	Experiment	Easy	Relatively Easy	Medium	Relatively Difficult	Difficult
Other						
gemini-2.0-flash	ICL	0.33333	0.22642	0.23762	0.26829	0.21341
gemini-2.0-flash	first-describe	0.32456	0.24528	0.21287	0.2439	0.25
gemini-2.0-flash	zero-shot	0.32456	0.32075	0.22277	0.26829	0.13415
gpt-4-turbo	ICL	0.18421	0.35849	0.21287	0.2439	0.14634
gpt-4-turbo	first-describe	0.17544	0.24528	0.19802	0.19512	0.21951
gpt-4-turbo	zero-shot	0.2193	0.26415	0.20297	0.19512	0.21341
gpt-4o	ICL	0.27193	0.28302	0.26238	0.2439	0.22561
gpt-4o	first-describe	0.25439	0.26415	0.21287	0.2439	0.28659
gpt-4o	zero-shot	0.28947	0.28302	0.26238	0.19512	0.25
gpt-4o-mini	ICL	0.18421	0.11321	0.19802	0.17073	0.19512
gpt-4o-mini	first-describe	0	0	0	0	0
gpt-4o-mini	zero-shot	0.16667	0.13208	0.23267	0.19512	0.21341
Human		0.58114	0.41528	0.44435	0.50975	0.53091

Table 5: Comparision of different model performance by different experiment for ”other” category

English Zero-Shot and Wronge-Image Prompt

Below, you can see multiple-choice questions (with answers).
Question: {mcq question}
Choices:
1)
2)
3)
4)
Answer:

English Without-Image Prompt

Below, you can see multiple-choice questions (with answers). If no image is provided, choose the best choice based on the available information.
Question: {mcq question}
Choices:
1)
2)
3)
4)
Answer:

E Difficulty Level & Trap Analysis

Model	Experiment	Easy	Relatively Easy	Medium	Relatively Difficult	Difficult
Social Science						
gemini-2.0-flash	ICL	0.90909	0.66667	0.69799	0.56522	0.42063
gemini-2.0-flash	first-describe	0.72727	0.41667	0.65101	0.54348	0.46032
gemini-2.0-flash	zero-shot	0.77273	0.41667	0.68456	0.54348	0.44444
gpt-4-turbo	ICL	0.72727	0.75	0.51007	0.34783	0.37302
gpt-4-turbo	first-describe	0.75	0.5	0.53435	0.38462	0.32759
gpt-4-turbo	zero-shot	0.72727	0.58333	0.4698	0.3913	0.30952
gpt-4o	ICL	0.63636	0.66667	0.57047	0.47826	0.35714
gpt-4o	first-describe	0.86364	0.58333	0.67114	0.52174	0.5
gpt-4o	zero-shot	0.72727	0.75	0.63087	0.54348	0.42063
gpt-4o-mini	ICL	0.45455	0.33333	0.32886	0.17391	0.19048
gpt-4o-mini	first-describe	0.375	0.28571	0.48889	0.29167	0.39394
gpt-4o-mini	zero-shot	0.77273	0.5	0.48322	0.43478	0.3254
Human		0.57772	0.48916	0.44785	0.33521	0.55833

Table 6: Comparision of different model performance by different experiment for ”Social Science” category

Model	Experiment	Easy	Relatively Easy	Medium	Relatively Difficult	Difficult
Natural Science						
gemini-2.0-flash	ICL	0.61856	0.52747	0.42128	0.40678	0.33787
gemini-2.0-flash	first-describe	0.68041	0.58242	0.53386	0.49831	0.49134
gemini-2.0-flash	zero-shot	0.66753	0.53297	0.46878	0.41695	0.40594
gpt-4-turbo	ICL	0.54897	0.46154	0.35972	0.35593	0.29208
gpt-4-turbo	first-describe	0.54381	0.3956	0.34125	0.32881	0.28094
gpt-4-turbo	zero-shot	0.55155	0.42308	0.36412	0.37966	0.30817
gpt-4o	ICL	0.61856	0.45604	0.39314	0.42034	0.32426
gpt-4o	first-describe	0.63918	0.58791	0.4635	0.46441	0.37129
gpt-4o	zero-shot	0.66753	0.53846	0.44943	0.44746	0.37252
gpt-4o-mini	ICL	0.35567	0.1978	0.19877	0.17627	0.15965
gpt-4o-mini	first-describe	0.49141	0.41333	0.34447	0.35849	0.2535
gpt-4o-mini	zero-shot	0.50773	0.37363	0.32454	0.31864	0.28713
Human		0.66489	0.50379	0.52244	0.43111	0.62153

Table 7: Comparision of different model performance by different experiment for ”Natural Science” category

Model	Experiment	Easy	Relatively Easy	Medium	Relatively Difficult	Difficult
Mathematics						
gemini-2.0-flash	ICL	0.41679	0.33824	0.32695	0.23077	0.25979
gemini-2.0-flash	first-describe	0.53973	0.57353	0.52531	0.53846	0.45077
gemini-2.0-flash	zero-shot	0.47676	0.48529	0.40903	0.50769	0.3618
gpt-4-turbo	ICL	0.27886	0.32353	0.25513	0.29231	0.20878
gpt-4-turbo	first-describe	0.28395	0.30882	0.23966	0.22222	0.1966
gpt-4-turbo	zero-shot	0.29985	0.36765	0.23598	0.30769	0.22894
gpt-4o	ICL	0.41079	0.48529	0.36389	0.4	0.33808
gpt-4o	first-describe	0.43028	0.42647	0.37893	0.44615	0.35469
gpt-4o	zero-shot	0.43478	0.41176	0.37346	0.4	0.33333
gpt-4o-mini	ICL	0.28036	0.27941	0.23598	0.27692	0.21827
gpt-4o-mini	first-describe	0.31049	0.375	0.24368	0.34286	0.27126
gpt-4o-mini	zero-shot	0.30885	0.29412	0.27497	0.35385	0.27639
Human		0.62299	0.45470	0.55884	0.37569	0.628374

Table 8: Comparision of different model performance by different experiment for ”Mathematics” category

Model	Experiment	Easy	Relatively Easy	Medium	Relatively Difficult	Difficult
Humanities						
gemini-2.0-flash	ICL	0.6	0.66667	0.8125	0.40625	0.36111
gemini-2.0-flash	first-describe	0.6	1	0.5625	0.6875	0.41667
gemini-2.0-flash	zero-shot	0.6	0.66667	0.75	0.53125	0.55556
gpt-4-turbo	ICL	0.4	0.33333	0.625	0.5	0.5
gpt-4-turbo	first-describe	0	0.5	0.6	0.48148	0.45
gpt-4-turbo	zero-shot	0.4	0.66667	0.5625	0.625	0.5
gpt-4o	ICL	0.4	0.66667	0.5	0.59375	0.47222
gpt-4o	first-describe	0.4	1	0.6875	0.71875	0.52778
gpt-4o	zero-shot	0.4	1	0.5625	0.59375	0.5
gpt-4o-mini	ICL	0.2	0.66667	0.4375	0.28125	0.16667
gpt-4o-mini	first-describe	0	0	0	0	0
gpt-4o-mini	zero-shot	0.2	0.33333	0.625	0.40625	0.27778
Human		0.57200	0.56333	0.45625	0.52031	0.60138

Table 9: Comparision of different model performance by different experiment for ”Humanities” category

Model	Experiment	% correct on Trap	% correct on None-Trap
gemini-2.0-flash	first-describe	0.4377	0.52437
gemini-2.0-flash	ICL	0.32068	0.39458
gemini-2.0-flash	zero-shot	0.3833	0.45042
gpt-4-turbo	first-describe	0.26381	0.30519
gpt-4-turbo	ICL	0.27135	0.32157
gpt-4-turbo	zero-shot	0.27641	0.32418
gpt-4o-mini	first-describe	0.26996	0.32428
gpt-4o-mini	ICL	0.20304	0.23137
gpt-4o-mini	zero-shot	0.27704	0.31989
gpt-4o	first-describe	0.37318	0.43716
gpt-4o	ICL	0.3365	0.4
gpt-4o	zero-shot	0.35863	0.43007
human		0.53485	0.56999

Table 10: Comparison of different models’ performance on trap and non-trap questions.