PerMed-MM: A Multimodal, Multi-Specialty Persian Medical Benchmark for Evaluating Vision Language Models

Ali Khoramfar, Mohammad Javad Dousti, and Heshaam Faili
Department of Electrical and Computer Engineering, College of Engineering,
University of Tehran, Tehran, Iran
{khoramfar, mjdousti, hfaili}@ut.ac.ir

Abstract

We present PerMed-MM, the first multimodal benchmark for Persian medical question answering. The dataset comprises 733 expert-authored multiple-choice questions from Iranian National Medical Board Exams, each paired with one to five clinically relevant images, spanning 46 medical specialties and diverse visual modalities. We evaluate five open-source and five proprietary vision language models, and find that reasoning supervision and domain-specific fine-tuning yield performance gains. Our cross-lingual analysis reveals significant unpredictability in translation-based pipelines, motivating the need for benchmarks that support direct, native-language evaluation. Additionally, domain- and modality-level analysis uncovers meaningful variation in model behavior often masked by aggregate metrics. PerMed-MM is publicly available on Hugging Face[1].

1 Introduction

Large language models (LLMs) have demonstrated unprecedented capabilities across a range of natural language processing (NLP) tasks. Yet, their unimodal nature has posed a fundamental limitation—restricting their ability to perceive and reason about the visual world. Recent advances in multimodal AI have addressed this gap through vision-language models (VLMs) and multimodal large language models (MLLMs), which enable joint processing of text and images within a unified architecture (Wu et al., 2023; Yin et al., 2024; Zhang et al., 2024).

These developments have sparked growing interest in applying VLMs to high-impact domains such as healthcare, where the ability to interpret both clinical text and medical images can support diagnostic reasoning, enhance medical education, and improve decision-making processes (Hartsock and Rasool, 2024; Meskó, 2023).

Standardized multiple-choice questions (MCQs) from official medical exams offer a valuable benchmark for objectively assessing the capabilities of LLMs, with recent studies demonstrating that advanced models can match and even surpass human performance thresholds (Liu et al., 2024).

With the emergence of VLMs, robust and interpretable evaluation frameworks have become important—particularly in clinical applications where in many real-world scenarios, effective decision-making depends on synthesizing multimodal inputs (Yan et al., 2023; Bazi et al., 2023). MCQs that are authored by medical experts and emphasize clinical reasoning over rote memorization provide a standardized, reliable means to track improvements in the performance of multimodal models.

Although several datasets for medical question answering (QA) have been introduced, current evaluation efforts remain limited in several key aspects. Most existing benchmarks focus on English (Alonso et al., 2024). Low-resource languages such as Persian remain largely overlooked, despite the pressing need for more equitable development in medical AI.

This underrepresentation is further compounded by the linguistic and resource-related challenges specific to Persian. The Persian language presents persistent challenges for LLMs due to its rich morphology, limited annotated data, and overall scarcity of resources—often resulting in lower performance compared to English (Arnett and Bergen, 2025; Abaskohi et al., 2024; Ryan et al., 2024).

To the best of our knowledge, there is currently no publicly available multimodal medical dataset in Persian. Existing resources have primarily centered on textual data, excluding visual modalities and thus constraining the development and assessment of VLMs in Persian clinical contexts.

[1] https://huggingface.co/datasets/universitytehran/PerMed-MM

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 232–241

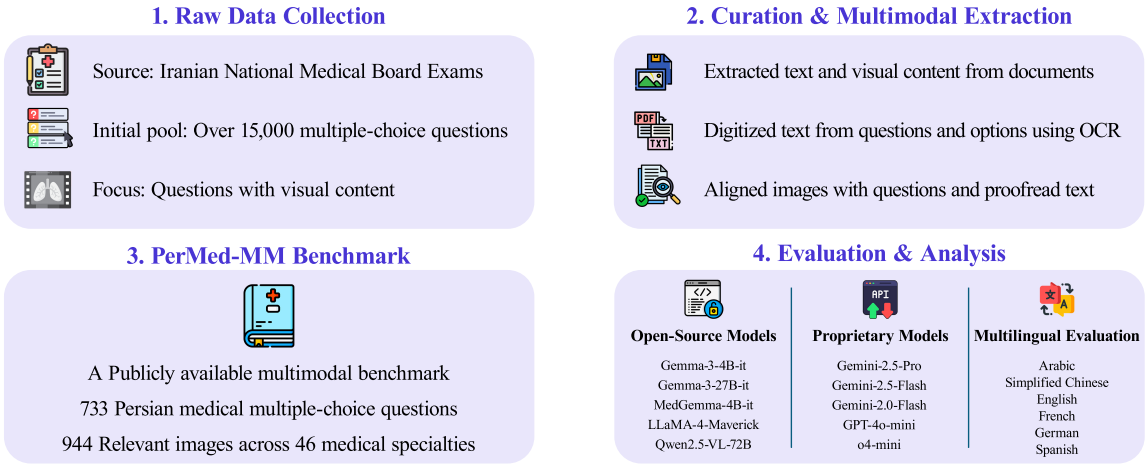


Figure 1: Overview of the workflow used to construct the PerMed-MM benchmark.

To address these gaps, we introduce PerMed-MM, a publicly available multimodal benchmark for Persian medical QA. Our main contributions are:

- **Multimodal coverage in a low-resource language:** PerMed-MM contains 733 Persian medical MCQs, each accompanied by one to five images, totaling 944 images, enabling multimodal evaluation in an underrepresented language.
- **Multi-specialty and modality diversity:** The dataset spans 46 medical specialties, encompassing a wide range of visual modalities, including radiographic images, histopathology slides, dermatologic photographs, ECG waveforms, and clinical charts or plots.
- **Baseline evaluation across diverse models:** We evaluate diverse open-source and proprietary models on the original Persian questions and their translations into multiple languages.

The rest of this paper is organized as follows. Section 2 reviews related work, Section 3 outlines our dataset and experimental setup, and Section 4 and Section 5 present our results and conclusions, respectively.

2 Related Work

Several benchmarks have been introduced to assess clinical reasoning in language models using MCQs from standardized medical exams. Early datasets such as MedQA (in English and Chinese) and MedMCQA (in English) introduced large-scale corpora targeting complex diagnostic tasks (Jin et al., 2021; Pal et al., 2022), while follow-up efforts like KorMedMCQA and MedQA-SWE extended this approach to less commonly represented languages such as Korean and Swedish (Kweon et al., 2024; Hertzberg and Lokrantz, 2024).

Multilingual and multimodal benchmarks have also been introduced for evaluating VLMs across domains. MMMU and EXAMS-V include image-text QA in multiple languages, covering a range of subjects with varied reasoning demands (Yue et al., 2024; Das et al., 2024).

In medicine, WorldMedQA-V provides clinical questions fully paired with diagnostic images in four languages (Matos et al., 2025), while HEAD-QA and MLEC-QA include medical questions in Spanish and Chinese, with a subset of questions linked to images (Vilares and Gómez-Rodríguez, 2019; Li et al., 2021).

Recent studies have assessed language models using questions from Persian national medical exams (Ebrahimian et al., 2023; Khorshidi et al., 2023; Zare et al., 2024). While these efforts provide insight into the role of language in model behavior, their evaluations have been limited to textual inputs, excluding visual information and offering no assessment of multimodal reasoning.

Several recent Persian medical QA resources provide complementary perspectives. PersianMedQA (Kalahroodi et al., 2025) offers 20,785 multiple-choice questions from national exams across 23 specialties, while PerMedCQA (Jamali et al., 2025) contains roughly 68,138 real-world, consumer-generated medical QAs from Persian medical forums—both limited to text. In contrast, PerMed-MM pairs 733 Persian MCQs with 944 relevant images spanning 46 specialties, enabling multimodal evaluation of vision–language reasoning in Persian medical context.

3 Methodology

Figure 1 provides an overview of our methodology, including data collection, multimodal extraction, benchmark construction, and model evaluation. We describe each component in detail below.

3.1 Data Collection

We compiled our dataset using publicly available data from the Iranian National Medical Specialty and Subspecialty Board exams held in 2021 and 2023. Initially, over 15,000 MCQs across 100 exams were manually reviewed to identify items containing visual content. For each selected item, images were extracted from the exam documents.

The exam PDFs were heterogeneous; some included extractable text, while others were scanned or produced errors when Persian text was parsed. To ensure consistent quality for PDFs with extractable text that still exhibited parsing issues, we applied both direct PDF text extraction and optical character recognition (OCR) using the Gemini-2.0-Flash model. The two outputs were cross-checked to detect and correct discrepancies, which substantially reduced noise and improved textual accuracy.

Subsequent manual reviews verified image–question alignment and confirmed answer correctness against the official exam keys. After cleaning, our final dataset comprises 733 multimodal MCQs, each paired with one to five images, spanning 46 medical specialties.

3.2 Baseline Model Evaluation

To evaluate baseline performance on PerMed-MM, we selected 10 leading models with diverse characteristics in terms of scale and reasoning capabilities. This set includes five open-source models—Gemma-3-4B-it, Gemma-3-27B-it (Team et al., 2025), MedGemma-4B-it (Sellergren et al., 2025), LLaMA-4-Maverick, and Qwen2.5-VL-72B-Instruct (Bai et al., 2025)—and five proprietary models accessed via API: Gemini-2.0-Flash, Gemini-2.5-Flash, Gemini-2.5-Pro (Comanici et al., 2025), GPT-4o-mini, and o4-mini. We included both reasoning-enabled and general-purpose models to assess how structured thinking impacts performance on multimodal clinical tasks.

Additionally, we incorporated MedGemma-4B-it, a domain-specialized variant of Gemma-3-4B-it, to directly examine the effect of medical fine-tuning on model behavior. Evaluations were conducted using a temperature of 0, and to facilitate a comparison of their problem-solving, we explicitly prompted all models—both reasoning and general-purpose—to generate a step-by-step chain of thought before arriving at a final answer.

Model	Snapshot / Update	Knowledge Cut-off
GPT-4o-mini	Jul 2024	Oct 2023
o4-mini	Apr 2025	Jun 2024
Gemini-2.0-Flash	Feb 2025	Aug 2024
Gemini-2.5-Flash	Jun 2025	Jan 2025
Gemini-2.5-Pro	Jun 2025	Jan 2025

Table 1: Configuration details for proprietary models accessed via commercial APIs.

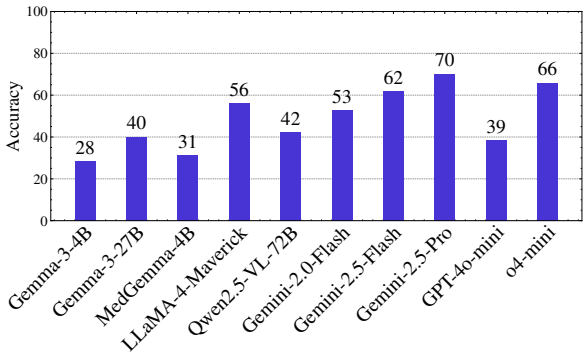


Figure 2: Baseline model accuracy on PerMed-MM.

To ensure consistent evaluation across commercial APIs, all models were run under similar configurations. Proprietary models were accessed through their official API endpoints, with versioning details and knowledge cut-offs summarized in Table 1. For the Gemini 2.5 series (Flash and Pro), we additionally specified `thinkingBudget = -1`, enabling adaptive reasoning depth that adjusts the internal thought length based on task complexity. Together, these settings helped maintain stable and comparable evaluation conditions across models.

3.3 Cross-Lingual Evaluation Setup

To investigate whether performance differences stemmed from language-related factors or medical knowledge, we conducted additional evaluations on questions translated into six languages: Arabic, Simplified Chinese, English, French, German, and Spanish. Translations were generated using Google Translate to enable a standardized and scalable comparison across languages. We selected three models—Gemma-3-4B-it, GPT-4o-mini, and Gemini-2.5-Flash—to represent three distinct categories of models.

Language	Gemini-2.5-Flash	Gemma-3-4B	GPT-4o-mini
Persian (Baseline)	62%	28%	39%
Arabic	56%	27%	36%
Chinese	58%	30%	38%
English	**60%**	**31%**	41%
French	59%	**31%**	**42%**
German	58%	29%	40%
Spanish	58%	29%	40%

Table 2: Accuracy of three models on MCQs translated into six languages, compared to the baseline.

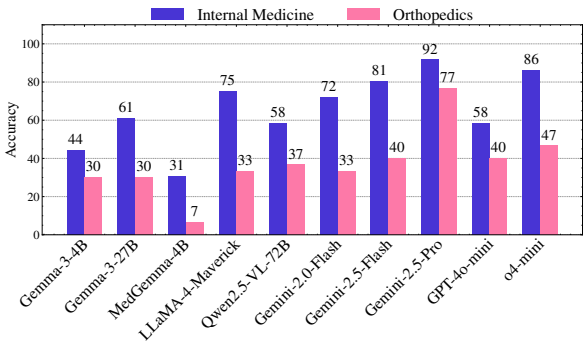


Figure 3: Models show strong performance in one domain but consistently underperform in another.

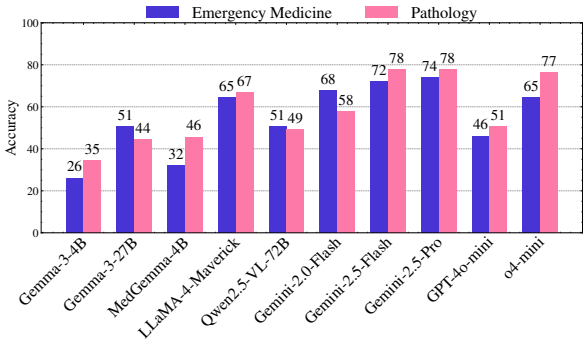


Figure 4: Models exhibit opposite performance trends across two clinical domains.

4 Results

4.1 Model Comparisons

Figure 2 presents the overall accuracy of the 10 models evaluated on PerMed-MM. Reasoning models—trained to generate structured chains of thought before producing final answers—consistently outperformed their counterparts across all evaluation settings. For instance, Gemini-2.5-Pro, explicitly trained for reasoning, achieved significantly higher accuracy than Gemini-2.0-Flash, which lacks such supervision—highlighting the effectiveness of combining advanced multimodal capabilities with structured reasoning. A similar pattern was observed when comparing o4-mini, a reasoning model, with GPT-4o-mini. These comparisons emphasized the value of models that are not only prompted to reason, but are inherently optimized to think before answering, leading to stronger performance on complex multimodal medical tasks. LLaMA-4-Maverick, with 400 billion parameters and a mixture-of-experts setup that activates about 17 billion per inference, yielded results consistent with expectations for its design. Meanwhile, Gemma-3-27B-it—despite being considerably smaller—achieved competitive performance, closely approaching that of the much larger Qwen2.5-VL-72B.

The performance difference between Gemma-3-4B-it and its domain-adapted counterpart, MedGemma-4B-it, underscores the value of specialized fine-tuning on medical data for enhancing multimodal understanding.

4.2 Multilingual Evaluation

Table 2 shows three models' accuracy on MCQs translated into six languages. Our multilingual evaluation investigates the assumption that a translate-then-inference pipeline improves performance for low-resource languages. While models generally showed higher accuracy on high-resource languages such as English, we found this pipeline approach could be detrimental. Specifically, Gemini-2.5-Flash achieved a higher accuracy of 62% on the original Persian questions compared to 60% on their English translations.

A deeper, question-level analysis revealed that this modest aggregate performance drop masked significant underlying instability. We found that 116 questions (15.8% of the data) had their correctness reversed between the two language settings. Crucially, in 64 of these cases, the model succeeded on the Persian (baseline) but failed on the English translation. To assess whether better translations could reduce this gap, we retranslated the questions using Gemini-2.5-Pro. With all other conditions unchanged, the number of cases where the model succeeded in Persian but failed in English dropped to 56. This finding underscores that for high-stakes domains, reliance on machine translation introduces unacceptable unpredictability, reinforcing the need to develop and evaluate models capable of direct reasoning in low-resource languages.

Model	With images	Text-only
Gemma-3-4B	28%	28%
LLaMA-4-Maverick	56%	53%
Gemini-2.0-Flash	53%	49%
Gemini-2.5-Flash	62%	55%
Gemini-2.5-Pro	70%	60%
GPT-4o-mini	39%	37%

Table 3: Effect of visual input on model accuracy (%) in PerMed-MM.

4.3 Domain-Specific Findings

Analysis of model performance across different subsets of the dataset revealed consistent variation in accuracy depending on the clinical domain. As illustrated in Figure 3, most models performed better on the knowledge-intensive questions of Internal Medicine, while struggling with the specialized visual reasoning required for Orthopedics. Notably, the domain-adapted MedGemma-4B-it experienced a significant performance drop in Orthopedics compared to its base model, Gemma-3-4B-it (30% to 7%). This suggests that the fine-tuning process may have inadvertently weakened its capabilities on certain multimodal tasks, a phenomenon that our benchmark can effectively identify.

Furthermore, a comparison between Emergency Medicine and Pathology revealed a crossover in model strengths. As shown in Figure 4, models like Gemini-2.0-Flash excelled in Emergency Medicine, while newer architectures like o4-mini showed a clear advantage in Pathology. This domain-specific evaluation enables comparison of model behavior beyond aggregate scores, offering a valuable tool to diagnose model-specific strengths, weaknesses, and areas for improvement. Full per-specialty accuracies for all models are reported in Table 7 in the Appendix.

4.4 Impact of Visual Modality

While PerMed-MM is designed as a multimodal benchmark, it is natural to ask whether models can answer questions from text alone or images provide any useful information. To investigate this, we conducted an ablation study in which a subset of models was evaluated in a text-only setting, where all images were removed and only the question text and multiple-choice options were provided. As shown in Table 3, removing images reduces accuracy for almost all models. This indicates that PerMed-MM questions cannot be reliably solved from textual cues alone.

5 Conclusion

This study introduces the first multimodal benchmark for evaluating clinical reasoning in Persian, built from 733 MCQs containing diverse visual content extracted from Iranian National Medical Board Exams. Unlike prior benchmarks that focused solely on text, our dataset enables assessment of models' reasoning across both textual and visual modalities, covering a wide range of medical specialties. Our evaluations across 10 vision-language models demonstrate that performance is significantly enhanced by reasoning-centric architectures. Crucially, our findings expose the unpredictability of translation-based pipelines for high-stakes tasks and uncover domain-specific weaknesses often masked by aggregate metrics.

Limitations

While PerMed-MM provides a new resource for medical evaluation in Persian, it has several limitations that should be considered.

First, our evaluation was shaped by the capabilities of the current model landscape. A key challenge was the limited Persian language support among most open-source medical models, which restricted the number of domain-specialized models we could assess.

Second, for cross-lingual evaluation, we rely on machine translations of the original Persian questions into other languages. However, since these translations have not been validated by native-speaking medical experts, they cannot be published as benchmarks for model evaluation.

Finally, the evaluation framework itself has inherent trade-offs. Although the multiple-choice question format offers both objective and scalable scoring, it inherently simplifies the nuanced and often ambiguous nature of real-world clinical reasoning.

Ethical Considerations

All data used in this study were obtained from publicly released Iranian National Medical Board Exams. The dataset does not contain any private, sensitive, or personally identifiable patient information, as the original questions and accompanying images are anonymized by design.

References

Amirhossein Abaskohi, Sara Baruni, Mostafa Masoudi, Nesa Abbasi, Mohammad Hadi Babalou, Ali Edalat, Sepehr Kamahi, Samin Mahdizadeh Sani, Nikoo Naghavian, Danial Namazifard, Pouya Sadeghi, and Yadollah Yaghoobzadeh. 2024. Benchmarking Large Language Models for Persian: A Preliminary Study Focusing on ChatGPT. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2189–2203. ELRA and ICCL.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. MedExpQA: Multilingual benchmarking of Large Language Models for Medical Question Answering. *Artificial Intelligence in Medicine*, 155:102938.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. 2023. Vision–Language Model for Visual Question Answering in Medical Imagery. *Bioengineering*, 10(3).

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *Preprint*, arXiv:2507.06261.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. EXAMS-V: A Multi-Discipline Multilingual Multimodal Exam Benchmark for Evaluating Vision Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7768–7791. Association for Computational Linguistics.

Manoochehr Ebrahimian, Behdad Behnam, Negin Ghayebi, and Elham Sobhrakhshankhah. 2023. ChatGPT in Iranian medical licensing examination: evaluating the diagnostic accuracy and decision-making capabilities of an AI-based model. *BMJ Health & Care Informatics*, 30(1):e100815.

Iryna Hartsock and Ghulam Rasool. 2024. Vision-language models for medical report generation and visual question answering: a review. *Frontiers in Artificial Intelligence*, Volume 7 - 2024.

Niclas Hertzberg and Anna Lokrantz. 2024. MedQA-SWE - a Clinical Question & Answer Dataset for Swedish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11178–11186. ELRA and ICCL.

Naghmeh Jamali, Milad Mohammadi, Danial Baledi, Zahra Rezvani, and Hesham Faili. 2025. PerMedCQA: Benchmarking large language models on medical consumer question answering in persian language. *Preprint*, arXiv:2505.18331.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14).

Mohammad Javad Ranjbar Kalahroodi, Amirhossein Sheikholselami, Sepehr Karimi, Sepideh Ranjbar Kalahroodi, Heshaam Faili, and Azadeh Shakery. 2025. PersianMedQA: Evaluating large language models on a persian-english bilingual medical question answering benchmark. *Preprint*, arXiv:2506.00250.

Hamid Khorshidi, Afshin Mohammadi, David M. Yousem, Jamileh Abolghasemi, Golnoosh Ansari, Mohammad Mirza-Aghazadeh-Attari, U Rajendra Acharya, and Ali Abbasian Ardakani. 2023. Application of ChatGPT in multilingual medical education: How does ChatGPT fare in 2023's Iranian residency entrance examination. *Informatics in Medicine Unlocked*, 41:101314.

Sunjun Kweon, Byungjin Choi, Gyouk Chu, Junyeong Song, Daeun Hyeon, Sujin Gan, Jueon Kim, Minkyu Kim, Rae Woong Park, and Edward Choi. 2024. KorMedMCQA: Multi-Choice Question Answering Benchmark for Korean Healthcare Professional Licensing Examinations. *Preprint*, arXiv:2403.01469.

Jing Li, Shangping Zhong, and Kaizhi Chen. 2021. MLEC-QA: A Chinese Multi-Choice Biomedical Question Answering Dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8862–8874. Association for Computational Linguistics.

Mingxin Liu, Tsuyoshi Okuhara, XinYi Chang, Ritsuko Shirabe, Yuriko Nishiie, Hiroko Okada, and Takahiro Kiuchi. 2024. Performance of ChatGPT Across Different Versions in Medical Licensing Examinations Worldwide: Systematic Review and Meta-Analysis. *J Med Internet Res*, 26:e60807.

João Matos, Shan Chen, Siena Kathleen V. Placino, Yingya Li, Juan Carlos Climent Pardo, Daphna Idan, Takeshi Tohyama, David Restrepo, Luis Filipe Nakayama, José María Millet Pascual-Leone, Guergana K Savova, Hugo Aerts, Leo Anthony Celi, An-Kwok Ian Wong, Danielle Bitterman, and Jack Gallifant. 2025. WorldMedQA-V: a multilingual, multimodal medical examination dataset for multimodal

language models evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7203–7216. Association for Computational Linguistics.

Bertalan Meskó. 2023. The Impact of Multimodal Large Language Models on Health Care's Future. *Journal of Medical Internet Research*, 25(1):e52865.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Michael J. Ryan, William Held, and Diyi Yang. 2024. Unintended Impacts of LLM Alignment on Global Representation. *Preprint*, arXiv:2402.15018.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. MedGemma Technical Report. *Preprint*, arXiv:2507.05201.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A Healthcare Dataset for Complex Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966. Association for Computational Linguistics.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023. Multimodal Large Language Models: A Survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256.

Zhiling Yan, Kai Zhang, Rong Zhou, Lifang He, Xiang Li, and Lichao Sun. 2023. Multimodal ChatGPT for Medical Applications: an Experimental Study of GPT-4V. *Preprint*, arXiv:2310.19061.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. *Preprint*, arXiv:2311.16502.

Soolmaz Zare, Soheil Vafaeian, Mitra Amini, Keyvan Farhadi, Mohammadreza Vali, and Ali Golestani. 2024. Comparing the performance of ChatGPT-3.5-Turbo, ChatGPT-4, and Google Bard with Iranian students in pre-internship comprehensive exams. *Scientific Reports*, 14(1):28456. Publisher: Nature Publishing Group.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-Language Models for Vision Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644.

A Benchmark Overview

A.1 Sample Questions

Figure 5 and Figure 6 show a representative question from the PerMed-MM benchmark, included here to illustrate the structure and visual format of our dataset. The original Persian version appears first, followed by its English translation used in cross-lingual evaluation.

Question (Persian)

خانم ۲۵ ساله با شکایت درد شدید هر دو هیپ با انتشار به پروگزیمال ران به اورژانس مراجعه کرده است. در سابقه مصرف منظم پودر چاق کننده را که از عطاری دریافت می‌کرده در یکسال گذشته دارد. در معاینه حرکات مفاصل هیپ در همه جهات محدود و دردناک است. استریا در پوست نواحی پهلوها رویت می‌شود. گرافی لگن را مشاهده می‌کنید.کدام تشخیص محتمل‌تر است؟

Associated Image

Multiple-Choice Options

1: اسپوندیلیت انکیلوزان
2: استئوآرتریت
3: نکروز آواسکولار
4: نقرس کاذب

Correct Answer: 3

Figure 5: Original Persian multiple-choice question from PerMed-MM.

Question (translated to English)

25-year-old woman presented to the emergency department with complaints of severe pain in both hips radiating to the proximal thigh. She has a history of regular use of fattening powder obtained from a herbalist for the past year. On examination, the hip joints are painful and limited in all directions. Striae are visible in the skin of the flanks. You can see the pelvic x-ray. Which diagnosis is most likely?

Associated Image

Multiple-Choice Options

1: Ankylosing spondylitis
2: Osteoarthritis
3: Avascular necrosis
4: Pseudogout

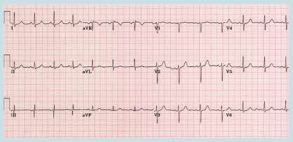
Correct Answer: 3

Figure 6: English translation of the same question for cross-lingual evaluation.

Multi-image Question (translated to English)

A 45-year-old man was referred to the emergency department with chest pain and shortness of breath. The pain had started 2 hours earlier and had not improved despite taking two sublingual nitroglycerin tablets. He first went to a clinic, where an ECG was obtained, and was then referred to the emergency department for further evaluation. The patient has a history of well-controlled hypertension. Currently, he reports no chest pain or other complaints. Vital signs and ECGs obtained at the clinic and the emergency department are as follows:
BP = 135/85 mmHg, PR = 98/min, RR = 16/min, SpO_2 = 96%
Clinic ECG:

Emergency department ECG:



Aspirin, heparin, and atorvastatin have been initiated. What is the next appropriate step in management?

Multiple-Choice Options

1: Immediate administration of reteplase
2: Emergent PCI
3: Serial ECG monitoring and troponin testing
4: CABG

Figure 7: Example of a Multi-image Question.

Figure 7 further illustrates a multi-image question where two complementary ECGs are provided for joint interpretation.

A.2 Image Count per Question

Images per Question	Count	Percentage
1	603	82.3%
2	83	11.3%
3	15	2.0%
4	30	4.1%
5	2	0.3%

Table 4: Distribution of questions in PerMed-MM based on the number of associated images.

Table 4 summarizes the number of images associated with each question in PerMed-MM. While most questions are paired with a single image, a notable subset includes multiple clinically relevant images—up to five in some cases—allowing evaluation of multimodal reasoning across varying image complexity.

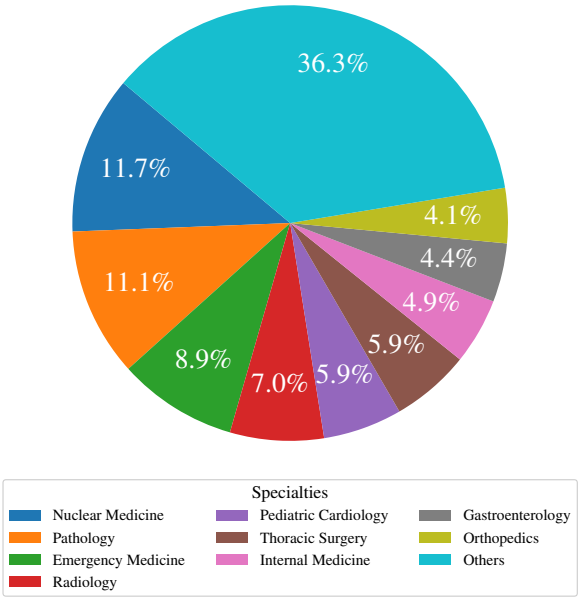


Figure 8: PerMed-MM question distribution across medical specialties.

A.3 Specialty Distribution

Figure 8 shows the percentage of PerMed-MM questions by specialty. Medical specialties with at least 30 questions in the benchmark are shown individually, while all others are grouped under the "Others" category.

B Prompt Templates

This section provides the prompt templates used in our experiments. The first prompt guided models to generate step-by-step clinical reasoning on multimodal multiple-choice questions.

B.1 Chain-of-Thought-Based Prompt

Evaluation Prompt for English Questions

You are a medical expert serving as an exam assistant. You will receive exactly one English multiple-choice medical question, four options numbered 1, 2, 3, and 4, and one to five images tagged as <Image 1>, <Image 2>, … <Image 5>.
First, provide a clear, step-by-step reasoning in English, referring explicitly to the question text and any images. Then, on the very last line, output ONLY the answer in this exact format:
answer: N
where N is one of 1, 2, 3, or 4. Do NOT include any extra text, punctuation, or explanation before or after this line.

While the structure was consistent across languages, the example shown here corresponds to the English version. The second prompt was used in the translation refinement experiment described at the end of Section 4.2.

B.2 Translation Prompt

Prompt for Translation Refinement

You are a professional translator with expertise in medical content. Your task is to translate a multiple-choice medical question from Persian to clear, natural English. The question includes four answer choices and may refer to one or more medical images. Ensure that all clinical terminology, context, and reasoning cues are preserved precisely in the English version. Avoid embellishment or paraphrasing.

This translation prompt was used in an auxiliary experiment discussed at the end of Section 4.2. The goal was to assess whether regenerating English translations using Gemini-2.5-Pro could reduce instability observed in the original translate-then-infer pipeline.

C Extended Analyses

Image Modality	Count	Percentage
Microscopic Pathology	175	18.5%
Charts, Diagrams & Tables	138	14.6%
X-ray	131	13.9%
CT	115	12.2%
Clinical / Gross Photography	97	10.3%
Ultrasound	76	8.1%
Electrophysiology (ECG / EEG)	73	7.7%
Nuclear Medicine	72	7.6%
MRI	46	4.9%
Endoscopy	21	2.2%

Table 5: Distribution of PerMed-MM images across modalities. Percentages are computed over all images in the benchmark.

C.1 Image Modality Analysis

As discussed in Section 4.4, PerMed-MM was explicitly designed to capture a diverse set of imaging modalities. Table 5 reports the distribution of image types in the benchmark, confirming that the dataset covers a broad spectrum of radiologic, endoscopic, microscopic, and clinically photographed content.

Image Modality	Gemma-3 4B	Gemma-3 27B	MedGemma 4B	LLaMA-4 Maverick	Qwen2.5-VL-72B	Gemini-2.0 Flash	Gemini-2.5 Flash	Gemini-2.5 Pro	GPT-4o mini	o4 mini
X-ray	28%	44%	30%	51%	42%	59%	58%	75%	46%	64%
CT	31%	39%	36%	57%	39%	50%	59%	68%	34%	71%
MRI	32%	39%	27%	51%	31%	46%	57%	78%	36%	67%
Nuclear Medicine	18%	34%	31%	48%	37%	41%	50%	55%	39%	57%
Ultrasound	25%	40%	29%	56%	43%	51%	58%	65%	34%	62%
Microscopic Pathology	35%	41%	38%	67%	46%	59%	78%	79%	49%	74%
Electrophysiology (ECG / EEG)	33%	37%	29%	51%	46%	54%	61%	68%	40%	61%
Endoscopy	38%	38%	26%	65%	38%	68%	73%	84%	30%	73%
Clinical / Gross Photography	23%	39%	29%	54%	43%	51%	57%	68%	30%	61%
Charts, Diagrams & Tables	29%	48%	34%	59%	48%	51%	65%	69%	36%	68%

Table 6: Accuracy (%) across image modalities for each evaluated model.

Specialty	Gemma-3 4B	Gemma-3 27B	MedGemma 4B	LLaMA-4 Maverick	Qwen2.5-VL-72B	Gemini-2.0 Flash	Gemini-2.5 Flash	Gemini-2.5 Pro	GPT-4o mini	o4 mini
Nuclear Medicine	22%	36%	30%	43%	34%	40%	52%	58%	34%	56%
Pathology	35%	44%	46%	67%	49%	58%	78%	78%	51%	77%
Emergency Medicine	26%	51%	32%	65%	51%	68%	72%	74%	46%	65%
Radiology	35%	39%	27%	37%	31%	51%	45%	59%	25%	59%
Pediatric Cardiology	19%	28%	37%	47%	35%	44%	58%	63%	37%	65%
Thoracic Surgery	35%	28%	37%	53%	42%	47%	56%	65%	35%	65%
Internal Medicine	44%	61%	31%	75%	58%	72%	81%	92%	58%	86%
Gastroenterology	34%	47%	31%	59%	44%	53%	72%	75%	34%	78%
Orthopedics	30%	30%	7%	33%	37%	33%	40%	77%	40%	47%

Table 7: Accuracy (%) by medical specialty and model for the nine most frequent specialties in PerMed-MM.

C.2 Accuracy by Image Modality × Model

Table 6 provides disaggregated model performance across imaging modalities in PerMed-MM. The results highlight that performance varies notably by visual modality.

C.3 Specialty-Level Performance

Table 7 reports accuracies for the nine most frequent specialties in PerMed-MM. These results complement Figures 3 and 4, showing that models can excel in certain domains while underperforming in others, even when overall accuracy is similar.

C.4 Cross-Lingual Consistency

Model	Persian	English
Gemma-3-4B	28%	31%
Gemma-3-27B	40%	41%
LLaMA-4-Maverick	56%	58%
Gemini-2.0-Flash	53%	49%
Gemini-2.5-Flash	62%	60%
Gemini-2.5-Pro	70%	69%
GPT-4o-mini	39%	41%
o4-mini	66%	64%

Table 8: Accuracy (%) of eight models on the original Persian questions and their English translations.

To verify the consistency of our findings, we extended the English evaluation to a broader set of models. As shown in Table 8, their accuracy patterns remained consistent with the initial subset, confirming that the overall cross-lingual trends observed in Section 4.2 hold across models.