WorldMedQA-V: a multilingual, multimodal medical examination dataset for multimodal language models evaluation

João Matos[1*], **Shan Chen**[2,3,4*], **Siena Placino**[5], **Yingya Li**[2,4], **Juan Carlos Climent Pardo**[2,3]
Daphna Idan[6], **Takeshi Tohyama**[7,9], **David Restrepo**[7], **Luis F. Nakayama**[7]
Jose M. M. Pascual-Leone[8], **Guergana Savova**[2,4], **Hugo Aerts**[2,3,10], **Leo A. Celi**[2,7,11]
A. Ian Wong[12], **Danielle S. Bitterman**[2,3,4], **Jack Gallifant**[2,3†]

[1]Oxford, [2]Harvard, [3]Mass General Brigham, [4]Boston Children's Hospital,
[5]St. Luke's Medical Center, [6]Ben-Gurion University of the Negev, [7]MIT, [8]Alcalá University,
[9]International University of Health and Welfare, [10]Maastricht University, [11]BIDMC, [12]Duke

Abstract

Multimodal/vision language models (VLMs) are increasingly being deployed in healthcare settings worldwide, necessitating robust benchmarks to ensure their safety, efficacy, and fairness. Multiple-choice question and answer (QA) datasets derived from national medical examinations have long served as valuable evaluation tools, but existing datasets are largely text-only and available in a limited subset of languages and countries. To address these challenges, we present WorldMedQA-V, an updated multilingual, multimodal benchmarking dataset designed to evaluate VLMs in healthcare. WorldMedQA-V includes 568 labeled multiple-choice QAs paired with 568 medical images from four countries (Brazil, Israel, Japan, and Spain), covering original languages and validated English translations by native clinicians, respectively. Baseline performance for common open- and closed-source models are provided in the local language and English translations, and with and without images provided to the model. The WorldMedQA-V benchmark aims to better match AI systems to the diverse healthcare environments in which they are deployed, fostering more equitable, effective, and representative applications.[1]

1 Introduction

Generative artificial intelligence (AI) models are increasingly being adopted in healthcare, highlighting the need for robust benchmarks to assess their safety, efficacy, and fairness (Thirunavukarasu et al., 2023; Clusmann et al., 2023; Abbasian et al., 2024; Wiggers, 2024).

One of the key evaluation tasks in Natural Language Processing (NLP) is Question Answering (QA)(Yu et al., 2024; Fan et al., 2023), which involves building systems that can automatically respond to human queries in natural language by combining language understanding with information retrieval (Jin et al., 2020). Multi-choice QA benchmarks have become essential not only for evaluating large language models (LLMs) but also for assessing vis language models (VLMs) in medicine (Liu et al., 2024).

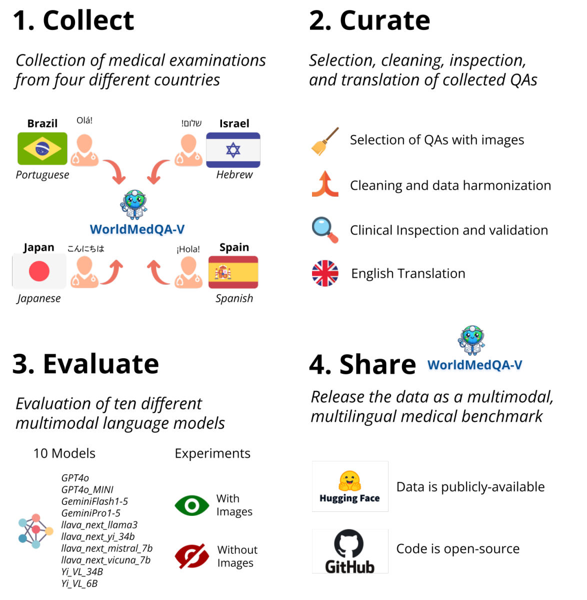


Figure 1: WorldMedQA-V dataset generation and evaluation workflows.

Recent research has explored the performance of LLMs in medical exams, with ChatGPT being the first AI system to pass the USMLE (Kung et al., 2023), prompting further studies (Gobira et al., 2023; Liu et al., 2024; Chen et al., 2024b). A recent review identified 45 studies on ChatGPT's performance in medical exams (Liu et al., 2024), but VLMs remain underexplored in medical tasks (Yan et al., 2023; Wu et al., 2023). Despite progress, current models face limitations such as context fragility, biases, and inconsistent multilingual per-

[*] Co-first authors: João Matos and Shan Chen

[†] Corresponding author: jgallifant@bwh.harvard.edu

[1] All code is accessible on https://github.com/WorldMedQA/V and the dataset on https://huggingface.com/datasets/WorldMedQA/V.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 7218–7231
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

formance (Gallifant et al., 2024; Zack et al., 2024; Chen et al., 2024a). There is also a need for more diverse datasets to ensure equitable AI evaluation in healthcare (Restrepo et al., 2024b). Key gaps include:

- **Real-world validity**: Studies reveal errors in existing medical QA datasets (Saab et al., 2024).
- **Linguistic diversity**: Many datasets lack language representation (Appendix Table 1) (Restrepo et al., 2024a,b; Ryan et al., 2024).
- **Imaging data**: Most medical QA benchmarks exclude multimodal data (Appendix Table 1)
- **Training data contamination**: Outdated datasets may overlap with LLM/VLM training corpora (Zhang et al., 2024a,b; Gallifant et al., 2024).

To address these issues, we introduce WorldMedQA-V, a multilingual, multimodal dataset for evaluating language and vision models. Key contributions include:

- **Multimodal medical exams from four countries**, supporting local languages and English.
- Previously **unseen** multimodal exam questions with **clinical validation** by medical professionals.
- Baseline **performance reporting of current state-of-the-art VLMs** across languages, including an evaluation of performance differentials between **local languages and English**.
- An investigation into the **impact of adding image data** to model performance and **stability across language** translations.

2 Related Work

Recent benchmarks like MMMU (Zhang et al., 2023b), MMMU-pro (Wang et al., 2024), EXAMS-V (Zhang et al., 2023a), and CulturalVQA (Wang et al., 2023) evaluate VLMs across multiple languages and disciplines, revealing notable performance gaps across linguistic and cultural contexts. Studies show that VLMs perform better in English, likely due to the predominance of English training data (Adam et al., 2023; Weidinger et al., 2021). These findings highlight the need for improving VLMs in diverse languages and cultural settings, especially in specialized domains. Moreover, all previous benchmarks hold limited amount of health/medical related questions that are not clinically verified.

Appendix A.1 Table 1 summarizes existing medical QA datasets by country. Six languages are covered, spanning seven administrative regions across three continents: Asia (China, India, South Korea, and Taiwan), Europe (Spain and Sweden), and North America (U.S.).

Medical datasets from these regions highlight challenges in LLMs' performance in healthcare. In Asia, notable datasets include those from China (Li et al., 2021a), Taiwan (Jin et al., 2020), South Korea (Kweon et al., 2024), and India (Pal et al., 2022). The MLEC-QA dataset from China, with 136,236 multiple-choice questions, is the largest. Despite LLMs being pre-trained on vast datasets, performance in this domain is hindered by limited diversity and quality of training data, especially for two-step reasoning and biomedical concepts (Li et al., 2021a). Similar trends are observed in Taiwan and South Korea, where English-pretrained models underperform on local medical exams. In Europe, datasets from Spain, Sweden, and Poland (the latter not publicly available) underscore the difficulties LLMs face, especially as question complexity increases (Vilares and Gómez-Rodríguez, 2019a). However, recent advancements saw models like *GPT3.5-Turbo* and *GPT4* pass the Swedish medical licensing exam (Hertzberg and Lokrantz, 2024a), while *GPT4-Turbo* slightly outperformed humans in Poland (Bean et al., 2024).

3 Methodology

Figure 1 shows the overall workflow of the study.

3.1 Data Collection

Our study uses medical exam data from Brazil, Israel, Japan, and Spain, consisting of multiple-choice questions from national licensing or specialization exams. Brazil's dataset includes 100 questions per exam from the 2011–16 and 2020–24 "Revalida" exams. Israel's dataset contains 150 questions from Phase A of the resident certification exam (2020–23). Japan's data comes from the 116th–118th National Medical Licensing Examinations (2022–24), while Spain's dataset includes questions from specialization exams (2019–23). Further details are provided in Appendix A.2.

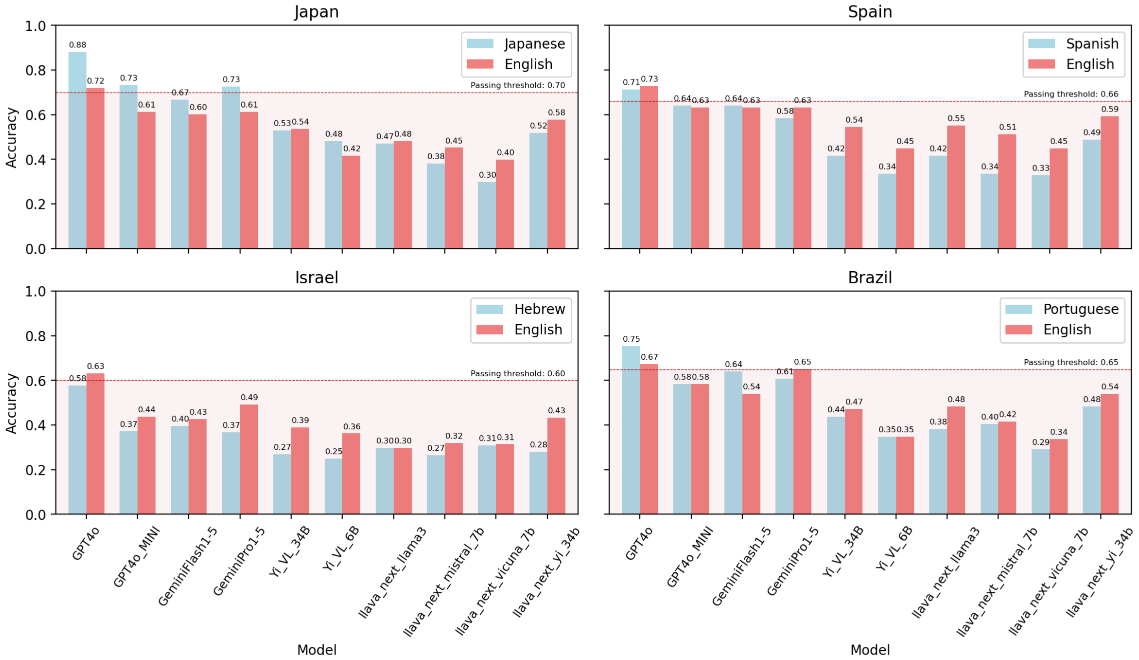


Figure 2: Accuracy in local language and English across models and countries. The red-shaded area highlights each country's exam passing threshold. Passing score is a proxy here since our dataset is a subset. Detailed results in Appendix A.6

3.2 Clinical Validation

A clinical validation process was carried out for all collected and translated data to ensure their quality and relevance. Native-speaking clinicians from each country validated the three key stages of the process — data extraction, translation, and final QA review.

3.3 Evaluation

Models: We included open- and closed-source models across a range of sizes: *GPT4o-2024-05-13*, *GPT4o-MINI-2024-07-18*, *GeminiFlash1-5 May*, *GeminiPro1-5 May*, *llava-next-llama3(8B)*, *llava-next-yi-34b*, *llava-next-mistral-7b*, *llava-next-vicuna-7b*, *Yi-VL-34B*, and *Yi-VL-6B*. All models were set to generate 512 tokens, with a temperature of 0 for reproducibility, and evaluated with Nvidia-GPU with CUDA > 12.0.

Experiments: The *VLMEvalKit* evaluation framework (Duan et al., 2024) was utilized to conduct experiments. We evaluated the ten models with and without image input, using accuracy as the metric. Cohen's kappa coefficients (Cohen, 1968) were computed to assess each model's reliability when answering the question in the original language versus the English translation.

4 Results and Discussion

Dataset: The complete WorldMedQA-V includes a total of 726 QAs and 850 images across four countries: Brazil, Israel, Japan, and Spain. Each QA is paired with at least one image, though some images appear in more than one question. After the exclusion of questions with multiple images or correct options, the final evaluation subset contains 568 QAs, each with a single associated image and correct option. Table 2, in Appendix A.3, provides a detailed summary of data distribution across countries and languages. Box 1 in Appendix A.4 shows an example from the Brazilian dataset.

VLMs' Performance: Figure 2 shows model performance across datasets in the local language and in English. Compared to the previously reported performance of *GPT4* on the USMLE, which is 90% (Brin et al., 2023), all models exhibit reduced performance when confronted with both image and text data. *GPT4o* emerged as the best-performing model. The only dataset for which *GPT4o* did not achieve a passing grade was the Israel dataset in Hebrew on which it achieved only 58%. Interestingly, *GPT4o* passed the English-translated version (63%) of the Israeli dataset. The other dataset in WorldMedQA-V with a non-Roman alphabet is the Japan dataset, on which *GPT4o* achieved an accuracy of 88%, exceed-

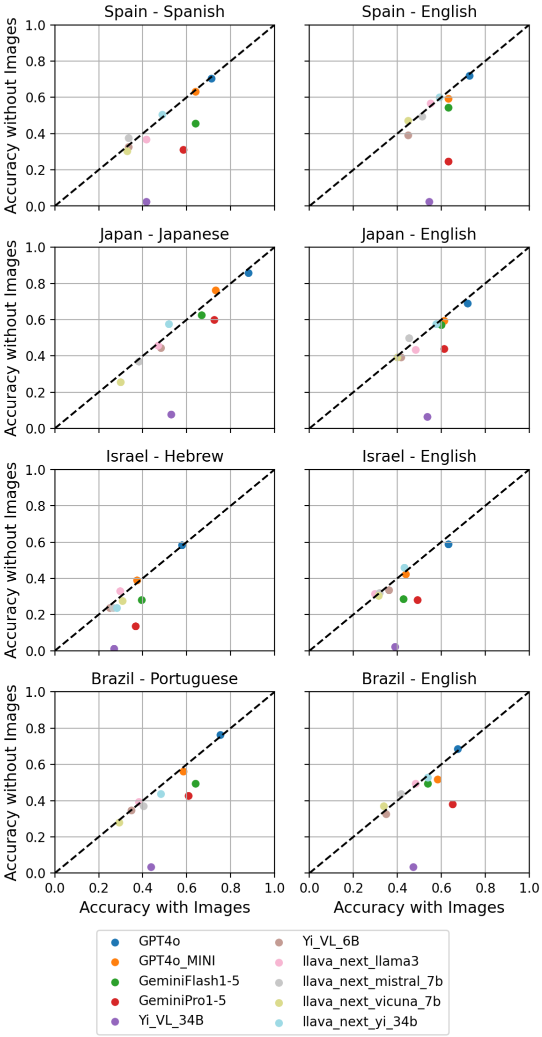


Figure 3: Accuracy across countries and languages with and without image input. See Appendix A.6 for details.

ing the 70% passing threshold. This may be because Japanese is better represented in pretraining datasets and has character overlap with Chinese characters/kanji. The underperformance in Hebrew, in contrast, could reflect Hebrew's lower representation in pretraining data, affecting the models' ability to understand the native language as effectively (Üstün et al., 2024). Models generally performed better on English-translated datasets, particularly for the Spain and Israel datasets. Moreover, the English-translated Israel dataset exhibits a somewhat lower overall performance when compared to other countries, which may indicate data variations that go beyond language. In the Brazilian subset, *GPT4o* scored 75% in Portuguese and 67% in English. Similarly, in the Japanese dataset, models such as *GPT4o*, *GPT4o-MINI*, *GeminiFlash1-5*, and *GeminiPro1-5* performed better in Japanese than in English, indicating strong language support for Japanese. The lowest accuracies were from the *llava-next* series, particularly on the Israel dataset, where several variants achieved nearly random accuracies in both Hebrew and English, ranging 24-46%. (Figure 2)

Accuracy with and without image input: Models performed better with image input. This trend was consistent across most datasets, particularly for models with lower baseline performance. However, the accuracy of the *GPT* models showed only minor variations — typically within 1-3% — regardless of whether the image was provided. Models from the *Gemini* family tended to be most sensitive to the exclusion of images, with improvements ranging from 4-27% when images were provided. The *Yi-VL* and *llava-next* models, which generally underperformed across the board, exhibited more stochastic variations in either direction depending on image input. Lastly, it is worth noting that the *Yi-VL-34b* model had almost no predictive power without images. (Figure 3)

Model consistency comparing English and local languages: Table 4 in Appendix A.5 compares model outputs in original languages to English translations using Cohen's kappa. *GPT4o* consistently achieved the highest agreement, particularly in the Brazil, Japan, and Spain datasets, with better performance in image-based settings. The highest kappa (84%) was observed in Spain's text-only setting, likely due to the model's high overall accuracy. Models like *GPT4o-MINI* and *GeminiFlash1-5* performed well in Brazil and Spain but lagged behind *GPT4o*. In contrast, *Yi-VL* showed lower agreement across countries, suggesting worse cross-language consistency. Notably, *GeminiPro1-5* showed an improvement in kappa, from 16.3% to 69.3%, when images were included in the Spanish set, demonstrating a substantial stabilizing effect of multimodal input. Overall, model cross-linguistic consistency improved with image data input.

5 Conclusion

In this work, we introduced `WorldMedQA-V`, a clinically validated, multilingual, and multimodal dataset containing medical QAs and images from Brazil, Israel, Japan, and Spain. We evaluated the performance of 10 vision-language models using both local languages and English translations, revealing performance disparities across languages and demonstrating how multimodal data can enhance accuracy. Despite improvements from image-based inputs, underrepresented languages like Hebrew proved particularly challenging. Throughout the data collection process, we engaged with 27 regions and collaborated closely with local physicians to ensure clinical and contextual relevance, ultimately focusing on the four regions that met our stringent criteria of high-quality translations and robust image-based multiple-choice questions. Each question underwent rigorous review by native-speaking clinicians to ensure linguistic precision and clinical validity, making WorldMedQA-V a gold-standard benchmark. Our methodology went beyond mere data aggregation, involving meticulous curation and validation to create a resource that is both scientifically robust and practically relevant. We adopted a single-correct-answer format consistent with standardized medical exams to simplify evaluation and ensure reproducibility, yet over 95% of evaluated models still failed to achieve passing performance—underscoring the inherent difficulty of integrating multilingual, multimodal information. Although our current dataset focuses on four regions, we remain committed to expanding its geographic scope in future iterations, particularly to include underrepresented areas such as parts of Africa and the Americas, thereby continuing to address critical gaps in the evaluation of healthcare-focused VLMs.

6 Limitations

While WorldMedQA-V represents a significant step toward creating a multilingual, multimodal benchmark for evaluating VLMs in healthcare, several limitations must be acknowledged.

First, the dataset, while carefully curated by trained physicians to ensure the validity of both questions and answers, remains relatively small. As we evaluated 568 multiple-choice questions and images, the sample size is limited in comparison to larger text-based benchmarks.

Second, the dataset only includes data from four countries: Brazil, Israel, Japan, and Spain, spanning three continents. This geographic limitation results in an underrepresentation of certain regions, particularly Africa, North and Central America, Oceania, and other parts of Asia.

Furthermore, although the benchmark introduces multimodal elements, it pairs only one image per question. Real-world clinical scenarios often involve multiple images from different time points or modalities, such as a sequence of X-rays, CT scans, and pathology slides. Another limitation is that text that is within images were not translated or adapted. English translations, although validated by native-speaking clinicians from each country, require further cross-validations, as these are typically nontrivial tasks.

Additionally, the lack of open-source multimodal medical language models restricts our ability to comprehensively evaluate and compare state-of-the-art health AI using WorldMedQA-V. Furthermore, since the models we tested were not originally trained for the medical domain, some LLMs (e.g., Gemini) refused to respond when no image was provided for certain questions, resulting in lower scores. When evaluating model performance against a passing threshold, a limitation is that our analysis relies on a limited set of multiple-choice questions with images, which may not provide consistent difficulty levels across different questions within the same exam.

Lastly, we set the underlying assumption that each question had only one correct answer, excluding cases where multiple correct answers were possible. This decision was made to simplify evaluation, but it may not reflect the inherent ambiguity and complexity found in both medical examinations and real-world medical scenarios where multiple treatment options or diagnoses can be valid.

Acknowledgments

Contributions: The authors contributed to this work as follows: JM, SC, and JG were responsible for the manuscript writing. JM, SC, JG, SP, JCCP, DI, DB, and AIW were involved in editing. The methods design was carried out by JM, SC, YL, AIW, DB, and JG . Data collection was conducted by JM, SP, DI, LFN, and DR. Data curation was done by JM, SC, SP, JCCP, DI, TT, DR, LFN, JMMPL, and JG. SC and YL performed the LVMs experiments. Visualizations were created by JM, SC, and DB. Data interpretation was a collective effort by all authors. Clinical validation was done by LFN (Brazil), DI (Israel), TT (Japan), and JMMPL (Spain). English translation validation was handled by LFN, DI, TT, JMMPL, JCCP, JM, and JG. Funding for the research was provided by GS, HA, LAC, AIW, and DB. Study validation was under DR, LFN, GS, HA, LAC, AIW, DB, and JG.

Funding: The authors acknowledge financial support from the Google PhD Fellowship (SC), the Woods Foundation (DB, SC), the National Institutes of Health (NIH) R01CA294033 (DB, SC, JG), U54CA274516-01A1 (DB, SC, JG), R01 EB017205 (LAC), U54 TW012043-01 (DS-I Africa; LAC), OT2OD032701 (Bridge2AI; LAC), U54MD012530 (REACH Equity; AIW), the National Science Foundation (NSF) ITEST 2148451 (LAC), the ASTRO-ACS Clinician Scientist Development Grant ASTRO-CSDG-24-1244514 (DB), and a Clarendon Scholarship from the University of Oxford (JM). The authors also thank Google Cloud for funding Gemini API inference costs.

References

Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, Olivier Gevaert, Li-Jia Li, Ramesh Jain, and Amir M. Rahmani. 2024. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *npj Digital Medicine*, 7(1):1–14.

Dillon C Adam et al. 2023. Generative ai for infectious diseases: An evaluation of chatgpt for medical translation. *PLOS Global Public Health*, 3(6):e0001673.

Israel Medicine Association. IMA - Israel Medicine Association.

The Israeli Medical Association. The interns' website | written examination questionnaires - stage a.

Andrew M. Bean, Karolina Korgul, Felix Krones, Robert McCraith, and Adam Mahdi. 2024. Exploring the landscape of large language models in medical question answering. *Preprint*, arXiv:2310.07225.

Dana Brin, Vera Sorin, Akhil Vaid, Ali Soroush, Benjamin S. Glicksberg, Alexander W. Charney, Girish Nadkarni, and Eyal Klang. 2023. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Scientific Reports*, 13:16492.

Shan Chen, Jack Gallifant, Mingye Gao, Pedro Moreira, Nikolaj Munch, Ajay Muthukkumar, Arvind Rajan, Jaya Kolluri, Amelia Fiske, Janna Hastings, Hugo Aerts, Brian Anthony, Leo Anthony Celi, William G. La Cava, and Danielle S. Bitterman. 2024a. Crosscare: Assessing the healthcare implications of pre-training data on language model bias. *Preprint*, arXiv:2405.05506.

Shan Chen, Yingya Li, Sheng Lu, Hoang Van, Hugo J W L Aerts, Guergana K Savova, and Danielle S Bitterman. 2024b. Evaluating the chatgpt family of models for biomedical reasoning and classification. *Journal of the American Medical Informatics Association*, 31(4):940–948.

Jan Clusmann, Fiona R. Kolbinger, Hannah Sophie Muti, Zunamys I. Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P. Veldhuizen, Sophia J. Wagner, and Jakob Nikolas Kather. 2023. The future landscape of large language models in medicine. *Communications Medicine*, 3(1):1–8.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira | Inep. Exame Nacional de Revalidação de Diplomas Médicos Expedidos por Instituições de Educação Superior Estrangeira (Revalida).

Ministerio de Sanidad. Formación Sanitaria Especializada.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. *Preprint*, arXiv:2407.11691.

Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. 2023. A bibliometric review of large language models research from 2017 to 2023. *Preprint*, arXiv:2304.02020.

Jack Gallifant, Shan Chen, Pedro Moreira, Nikolaj Munch, Mingye Gao, Jackson Pond, Leo Anthony Celi, Hugo Aerts, Thomas Hartvigsen, and Danielle Bitterman. 2024. Language models are surprisingly fragile to drug names in biomedical benchmarks. *Preprint*, arXiv:2406.12066.

Mauro Gobira, Luis Filipe Nakayama, Rodrigo Moreira, Eric Andrade, Caio Vinicius Saito Regatieri, and Rubens Belfort Jr. 2023. Performance of ChatGPT-4 in answering questions from the Brazilian National Examination for Medical Degree Revalidation. *Revista da Associação Médica Brasileira*, 69(10):e20230848.

Niclas Hertzberg and Anna Lokrantz. 2024a. 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, Torino, Italia. ELRA and ICCL.

Niclas Hertzberg and Anna Lokrantz. 2024b. 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, Torino, Italia. ELRA and ICCL.

Japanese Ministry of Health Labour and Welfare 2022. The 116th National Medical Examination Questions and Answers.

Japanese Ministry of Health Labour and Welfare 2023. The 117th National Medical Examination Questions and Answers.

Japanese Ministry of Health Labour and Welfare 2024. The 118th National Medical Examination Questions and Answers.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *CoRR*, abs/2009.13081.

T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, and V. Tseng. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198.

Sunjun Kweon, Byungjin Choi, Minkyu Kim, Rae Woong Park, and Edward Choi. 2024. Kormedmcqa: Multi-choice question answering benchmark for korean healthcare professional licensing examinations. *arXiv*.

Jing Li, Shangping Zhong, and Kaizhi Chen. 2021a. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jing Li, Shangping Zhong, and Kaizhi Chen. 2021b. 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

M. Liu, T. Okuhara, X. Chang, R. Shirabe, Y. Nishiie, H. Okada, and T. Kiuchi. 2024. Performance of chatgpt across different versions in medical licensing examinations worldwide: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 26:e60807.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering.

David Restrepo, Luis Filipe Nakayama, Robyn Gayle Dychiao, Chenwei Wu, Liam G. McCoy, Jose Carlo Artiaga, Marisa Cobanaj, João Matos, Jack Gallifant, Danielle S. Bitterman, Vincenz Ferrer, Yindalon Aphinyanaphongs, and Leo Anthony Celi. 2024a. Seeing beyond borders: Evaluating llms in multilingual ophthalmological question answering. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 565–566.

David Restrepo, Chenwei Wu, Constanza Vásquez-Venegas, João Matos, Jack Gallifant, Leo Anthony Celi, Danielle S. Bitterman, and Luis Filipe Nakayama. 2024b. Analyzing Diversity in Healthcare LLM Research: A Scientometric Perspective.

Michael J. Ryan, William Held, and Diyi Yang. 2024. Unintended Impacts of LLM Alignment on Global Representation. *arXiv preprint*. ArXiv:2402.15018 [cs].

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, Si-Wai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of gemini models in medicine. *arXiv*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940.

David Vilares and Carlos Gómez-Rodríguez. 2019a. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

David Vilares and Carlos Gómez-Rodríguez. 2019b. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Yiyi Wang et al. 2023. Culturalvqa: A new frontier in vision and language understanding. In *Proceedings of CVPR*.

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

Laura Weidinger et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Kyle Wiggers. 2024. Hugging Face releases a benchmark for testing generative AI on health tasks.

Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Can gpt-4v(ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *Preprint*, arXiv:2310.09909.

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.

Huizi Yu, Lizhou Fan, Lingyao Li, Jiayan Zhou, Zihui Ma, Lu Xian, Wenyue Hua, Sijia He, Mingyu Jin, Yongfeng Zhang, Ashvin Gandhi, and Xin Ma. 2024. Large language models in biomedical and health informatics: A bibliometric review. *Preprint*, arXiv:2403.16303.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, Atul J Butte, and Emily Alsentzer. 2024. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

Andy K. Zhang, Kevin Klyman, Yifan Mai, Yoav Levine, Yian Zhang, Rishi Bommasani, and Percy Liang. 2024a. Language model developers should report train-test overlap. *arXiv preprint*. ArXiv:2410.08385 [cs].

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. 2024b. A Careful Examination of Large Language Model Performance on Grade School Arithmetic. *arXiv preprint*. ArXiv:2405.00332 [cs].

Jieyi Zhang et al. 2023a. Exams-v: A multi-discipline multi-lingual multi-modal exam benchmark for evaluating vision language models. *arXiv preprint arXiv:2308.03463*.

Xiang Zhang, Junyang Yang, Jianwei Zhang, et al. 2023b. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of NeurIPS*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A Appendix

A.1 Existing publicly-available medical examination QA dataset per country

Table 1: Summary of existing open-source Medical QA Datasets by Country.

Country	Dataset	#QA	Language(s)	Modalities	Source	Years
China	MedQA (Jin et al., 2020)	34,251	Simplified Chinese	Text	MCMLE, Mainland China Medical Licensing Examination	Not Clear
China	MLEC-QA (Li et al., 2021b)	136,236	Simplified Chinese	Text, Images	National Medical Licensing Examination (NMLEC)	Not Clear
India	MedMCQA (Pal et al., 2022)	193,155	English	Text	AIIMS PG, NEET PG	1991-2022
Spain	Head-QA (Vilares and Gómez-Rodríguez, 2019b)	6,765	Spanish and English	Text, Images	Ministerio de Sanidad, Consumo y Bienestar Social	2013–2017
Republic of Korea	KorMedMCQA (Kweon et al., 2024)	5,345	Korean and English	Text	Korea Health Personnel Licensing Examination Institute	2012–2023
Sweden	MedQA-SWE (Hertzberg and Lokrantz, 2024b)	3,180	Swedish	Text	National Board of Health and Welfare, Umeå University	2016–2023
Taiwan	MedQA (Jin et al., 2020)	14,123	Traditional Chinese	Text	TWMLE, Taiwan Medical Licensing Examination	Not Clear
United States	MedQA (Jin et al., 2020)	12,723	English	Text	USMLE, United States Medical Licensing Examination	Not Clear

A.2 Details on collected data per country

A.2.1 Brazil

The examination data were collected from the "Revalida" examinations, which are publicly available on the Brazilian government's official website. The "Revalida" exam, administered by the National Institute of Educational Research and Studies (INEP), supports the process of diploma revalidation for doctors who graduated abroad and wish to practice in Brazil. The exams consist of two sections: 100 multiple-choice questions (20 in each of the following areas: Internal Medicine, Surgery, Pediatrics, Preventive Medicine, and Gynecology and Obstetrics) and open-ended questions. For this work, only the multiple-choice section was included. Data from the years 2011–2016 and 2020–2024 were used, encompassing all publicly available years at the time of this study (de Estudos e Pesquisas Educacionais Anísio Teixeira | Inep).

A.2.2 Israel

The Israeli subset consists of questions from seven medical specialties: Internal Medicine, Clinical Microbiology, Neurology, Oncology, Ophthalmology, Urology, and Public Health. These questions are drawn from Phase A of the two-phase examination process that residents in Israel must complete during their training. Phase A is a written exam held annually, comprising approximately 150 questions. We included questions from tests administered between 2020 and 2023, with full versions of these exams publicly available on the Israel Medical Association's website (Association; Association).

The questions are categorized into three main types: preclinical cases, clinical cases, and questions based on scientific articles. Preclinical cases focus on foundational scientific knowledge, while clinical cases present patient background information followed by clinical questions related to the patient's medical conditions. Questions derived from scientific articles involve analysis of graphs, figures, and study results, which are particularly prevalent in the public health exam. Some questions also include visual aids, such as diagnostic images, laboratory slides (e.g., blood smear slides in Clinical Microbiology), and other data specific to patients' clinical presentations.

A.2.3 Japan

Japanese questions were sourced from past examinations that were published on the website of the Ministry of Health, Labour and Welfare. We included the 116th, 117th, and 118th National Medical Licensing Examination (Japanese Ministry of Health Labour and Welfare 2022; Japanese Ministry of Health Labour and Welfare 2023; Japanese Ministry of Health Labour and Welfare 2024), which corresponded to 2022-2024.

A.2.4 Spain

The examination data were sourced from the annual exams organized by the Ministerio de Sanidad, Consumo y Bienestar Social (Spanish Ministry of Health, Consumer Affairs, and Social Welfare) (de Sanidad). These exams are part of the competitive selection process for specialized medical positions in Spain's public healthcare system. Eligibility for participation requires candidates to possess a bachelor's degree in medicine (6 years of study) and typically prepare for a year or more, given the limited number of vacancies. The exams play a critical role in ranking candidates, who are able to select their specialization and hospital placement only based on their exam performance. For this study, only data from the years 2019–2023 were used to avoid overlap with the existing Head-QA dataset (Vilares and Gómez-Rodríguez, 2019b).

A.3 Detailed Data Statistics

Table 2: WorldMedQA's data across countries and languages. In the curated dataset, each QA was associated with at least one image. Some images were present in more than one question. In the final subset for evaluation (rightmost column), each question had a single image and the correct option associated with it, resulting in fewer samples. The number of answer options per question (fourth column) refers to the original number of choices in the multiple-choice format (e.g., A-D for four options or A-E for five options). However, all questions after preprocessing results in 4 options only. In cases where this varies, such as in Brazil, the value represents a weighted average across questions. The total number of QAs and images does not immediately add up due to some questions sharing images or having multiple associated options.

Country	**Language**	**Years**	**Option/QA**	**QAs, n (%)**	**Images, n (%)**	**Final, n (%)**
Brazil	Portuguese	2011-2024	4.27	93 (12.8%)	94 (11.1%)	89 (15.7%)
Israel	Hebrew	2020-2023	4.00	200 (27.6%)	184 (21.6%)	186 (32.7%)
Japan	Japanese	2022-2024	5.00	306 (42.1%)	445 (52.4%)	168 (29.6%)
Spain	Spanish	2019-2023	4.00	127 (17.5%)	127 (14.9%)	125 (22.0%)
Total	4 Languages	2011-2024	4.00	726 (100%)	850 (100%)	568 (100%)

A.4 Example QA from the Brazilian dataset

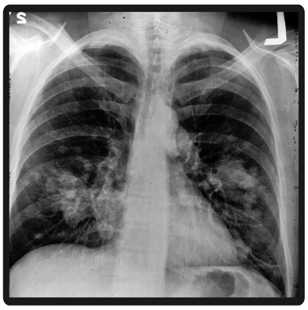
Box 1. Example multimodal QA from the Brazilian subset

Original (Portuguese)

Um paciente do sexo masculino, 55 anos de idade, tabagista 60 maços/ano, com tosse crônica há mais de 10 anos, relata que há cerca de três meses observou a presença de sangue na secreção eliminada com a tosse. Refere ainda perda de cerca de 15% do peso habitual nesse mesmo período, anorexia, adinamia e sudorese noturna. A radiografia de tórax realizada por ocasião da consulta é mostrada abaixo. Qual a hipótese diagnóstica mais provável nesse caso?

A) Aspergilose pulmonar.
B) Carcinoma pulmonar.
C) Tuberculose cavitária.
D) Bronquiectasia com infecção.
E) Doença pulmonar obstrutiva crônica.

Image



Translation (English)

A 55-year-old male patient, with a smoking history of 60 pack-years, has had a chronic cough for over 10 years. He reports that about three months ago, he noticed the presence of blood in the sputum. He also mentions a weight loss of about 15% of his usual weight during the same period, anorexia, weakness, and night sweats. The chest X-ray taken at the time of the consultation is shown below. What is the most likely diagnostic hypothesis in this case?

A) Pulmonary aspergillosis.
B) Lung carcinoma.
C) Cavitary tuberculosis.
D) Bronchiectasis with infection.
E) Chronic obstructive pulmonary disease.

A.5 Model output consistency across countries and test setting

Table 3: Cohen's Kappa reflecting agreement between languages for the same models, countries, and testing setting. Values in **bold** highlight the model with highest kappa per country and testing mode. The two studied settings were text-only (T. only) and text and image (T. & I.).

Country	**Brazil**		**Israel**		**Japan**		**Spain**	
Model ↓ Setting →	**T. only**	**T. & I.**	**T. only**	**T. & I.**	**T. only**	**T. & I.**	**T. only**	**T. & I.**
GPT4o	**0.684**	**0.743**	**0.619**	0.654	**0.683**	**0.618**	**0.840**	**0.829**
GPT4o-MINI	0.642	0.655	0.458	0.603	0.525	0.554	0.715	0.809
GeminiFlash1-5	0.612	0.536	0.533	**0.655**	0.591	0.521	0.594	0.767
GeminiPro1-5	0.389	0.469	0.184	0.416	0.351	0.490	0.163	0.693
Yi-VL-34B	0.020	0.438	0.111	0.309	0.033	0.393	0.030	0.507
Yi-VL-6B	0.427	0.320	0.150	0.204	0.240	0.348	0.269	0.251
llava-next-llama3	0.429	0.441	0.401	0.380	0.269	0.264	0.435	0.433
llava-next-mistral-7b	0.498	0.310	0.148	0.243	0.153	0.234	0.466	0.348
llava-next-vicuna-7b	0.385	0.491	0.167	0.281	0.091	0.185	0.310	0.279
llava-next-yi-34b	0.592	0.635	0.208	0.223	0.393	0.373	0.594	0.488

A.6 Detailed performance with and without images

Table 4: Accuracy comparison across countries and **original** languages (Portuguese, Hebrew, Japanese, and Spanish) for each model. The two studied settings were text-only (T. only) and text and image (T. & I.). Each cell represents the performance of each model in its native language dataset, highlighting how the presence or absence of images affects accuracy.

Country	**Brazil**		**Israel**		**Japan**		**Spain**	
Model ↓ Setting →	**T. only**	**T. & I.**	**T. only**	**T. & I.**	**T. only**	**T. & I.**	**T. only**	**T. & I.**
GPT4o	0.764	0.753	0.584	0.578	0.857	0.881	0.704	0.712
GPT4o-MINI	0.562	0.584	0.389	0.373	0.762	0.732	0.632	0.640
GeminiFlash1-5	0.494	0.640	0.281	0.395	0.625	0.667	0.456	0.640
GeminiPro1-5	0.427	0.607	0.135	0.368	0.601	0.726	0.312	0.584
Yi-VL-34B	0.034	0.438	0.011	0.270	0.077	0.530	0.024	0.416
Yi-VL-6B	0.348	0.348	0.238	0.249	0.446	0.482	0.328	0.336
llava-next-llama3	0.393	0.382	0.330	0.297	0.458	0.470	0.368	0.416
llava-next-mistral-7b	0.371	0.404	0.238	0.265	0.369	0.381	0.376	0.336
llava-next-vicuna-7b	0.281	0.292	0.276	0.308	0.256	0.298	0.304	0.328
llava-next-yi-34b	0.438	0.483	0.238	0.281	0.577	0.518	0.504	0.488

Table 5: Accuracy comparison across countries and **English-translated** datasets for each model. The two studied settings were text-only (T. only) and text and image (T. & I.). Each cell represents the performance of each model after translation, highlighting how the presence or absence of images affects accuracy.

Country	**Brazil**		**Israel**		**Japan**		**Spain**	
Model ↓ Setting →	**T. only**	**T. & I.**	**T. only**	**T. & I.**	**T. only**	**T. & I.**	**T. only**	**T. & I.**
GPT4o	0.685	0.674	0.589	0.632	0.690	0.720	0.720	0.728
GPT4o-MINI	0.517	0.584	0.422	0.438	0.595	0.613	0.592	0.632
GeminiFlash1-5	0.494	0.539	0.286	0.427	0.571	0.601	0.544	0.632
GeminiPro1-5	0.382	0.652	0.281	0.492	0.440	0.613	0.248	0.632
Yi-VL-34B	0.034	0.472	0.022	0.389	0.065	0.536	0.024	0.544
Yi-VL-6B	0.326	0.348	0.335	0.362	0.393	0.417	0.392	0.448
llava-next-llama3	0.494	0.483	0.314	0.297	0.435	0.482	0.568	0.552
llava-next-mistral-7b	0.438	0.416	0.319	0.319	0.500	0.452	0.496	0.512
llava-next-vicuna-7b	0.371	0.337	0.303	0.314	0.393	0.399	0.472	0.448
llava-next-yi-34b	0.528	0.539	0.459	0.432	0.577	0.577	0.600	0.592