npj | digital medicine

Brief communication

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01837-2

Benchmarking vision-language models for diagnostics in emergency and critical care settings

Christoph F. Kurz[1] ✉, Tatiana Merzhevich[2], Bjoern M. Eskofier[2,3], Jakob Nikolas Kather[4,5,6] & Benjamin Gmeiner[1]

The applicability of vision-language models (VLMs) for acute care in emergency and intensive care units remains underexplored. Using a multimodal dataset of diagnostic questions involving medical images and clinical context, we benchmarked several small open-source VLMs against GPT-4o. While open models demonstrated limited diagnostic accuracy (up to 40.4%), GPT-4o significantly outperformed them (68.1%). Findings highlight the need for specialized training and optimization to improve open-source VLMs for acute care applications.

Artificial intelligence (AI) is rapidly transforming healthcare, with increasing applications in diagnostics, clinical workflows, and patient management[1]. This potential is particularly evident in critical care environments such as emergency departments (EDs) and intensive care units (ICUs), where clinicians must synthesize imaging, text-based information, and real-time patient data under significant time constraints[2]. Vision-language models (VLMs), which merge natural language processing and computer vision, offer new possibilities for integrating multimodal data streams[3]. They promise to support decision-making by identifying diagnostic patterns from imaging studies, extracting and interpreting clinical notes, and combining disparate inputs within critical care environments[4–9]. For example, Chua et al.[10] showed that combining imaging features with clinical text improved early detection of sepsis in ED settings. While proprietary models like GPT-4 have demonstrated impressive capabilities, small open-source VLMs offer significant advantages in terms of scalability, data privacy, and cost, especially for resource-constrained and underserved settings[11].

Yet despite their growing availability, these models have not been systematically evaluated in image-based diagnostic tasks for acute care. For these reasons, we benchmarked several open-source VLMs chosen for their smaller computational footprints and suitability in clinical settings[12]. Delayed or inaccurate diagnoses in ED/ICU workflows can be catastrophic, underscoring the urgent need for robust diagnostic support. By integrating imaging and clinical data, VLMs may enhance diagnostic speed and accuracy for critical conditions such as stroke, myocardial infarction, sepsis, and trauma. As interest in AI-based diagnostic tools for acute care continues to grow, the New England Journal of Medicine (NEJM) Image Challenge dataset provides a valuable benchmark for assessing model performance on diverse, real-world diagnostic tasks with clearly defined ground-truth answers. Its rich collection of acute-care cases makes it particularly relevant for evaluating AI's potential to streamline high-stakes clinical decision-making.

In this study, we evaluated several open-source VLMs and compared them to GPT-4o using the NEJM Image Challenge dataset of over 1000 diagnostic questions, each linked to clinical images. The integration of clinical images, case descriptions, multiple-choice answers, and human responses allowed for a systematic assessment of model performance across varying complexity levels (Fig. 1)[12].

Our findings revealed that current small open-source VLMs lag significantly behind GPT-4o in diagnostic tasks. While GPT-4o achieved an accuracy of 68.1% and exceeding the average performance of human responders, open-source VLMs yielded poor accuracy, ranging from below random guessing (<20%) to a modest 40.4% for the largest, best-performing model (Fig. 2). Smaller VLMs designed for resource-constrained deployment, such as DeepSeek VL2 Tiny or Smol 500 M, were particularly inadequate, with some models performing below random guessing. Performance differences across difficulty levels were relatively minor, indicating that model accuracy is consistent across easier and more challenging diagnostic cases (Fig. 3). Notably, cases in the NEJM Image Challenge frequently overlapped with conditions that demand urgent intervention in

[1]Novartis Pharma GmbH, Nuremberg, Germany. [2]Machine Learning and Data Analytics (MaD) lab, Department of Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander Universität, Erlangen-Nürnberg (FAU), Erlangen, Germany. [3]Institute of AI for Health, Helmholtz Zentrum München, Munich, Germany. [4]Else Kroener Fresenius Center for Digital Health, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany. [5]Department of Medicine I, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany. [6]Medical Oncology, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany.
✉e-mail: christoph.kurz@novartis.com

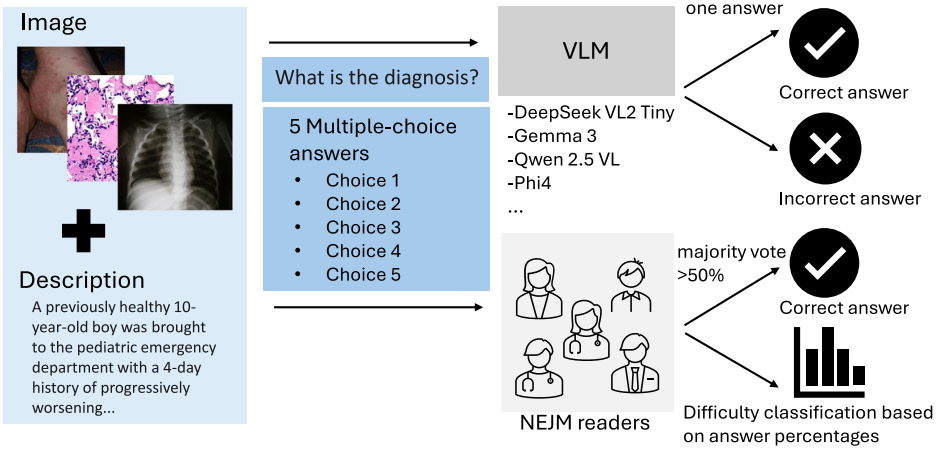

Fig. 1 | Overview of the benchmarking process for evaluating VLMs on the NEJM Image Challenge. VLMs analyze medical images and descriptions to select the correct diagnosis from multiple-choice answers, compared against expert readers' consensus. Medical images shown are sourced from Wikimedia Commons.

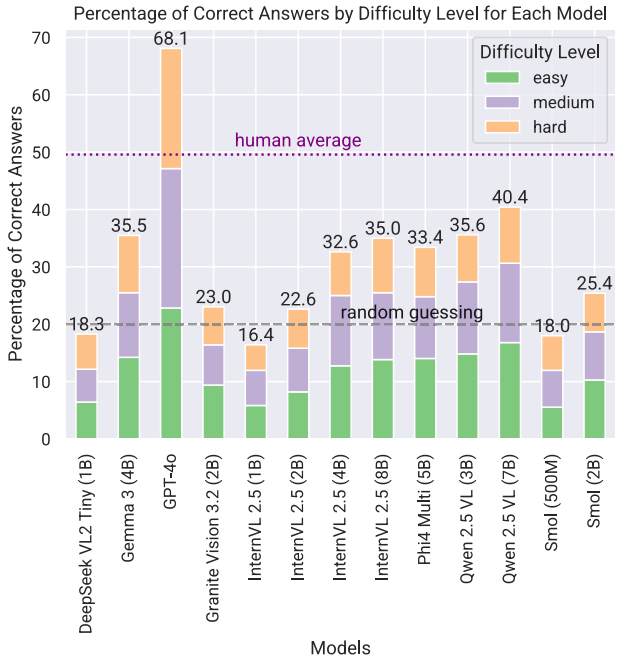

Fig. 2 | Percentage of correct answers for each model by difficulty level. Each bar is segmented to show results for easy (green), medium (purple), and hard (orange) questions. The total height of each bar represents the overall percentage of correct answers for that model. The horizontal dashed line indicates the random guessing threshold. Because there were five multi-choice answers in each challenge, we defined random guessing at 20% accuracy. The horizontal dotted line represents the average performance of human responders to the challenge. DeepSeek VL2 Tiny (1B), InternVL 2.5 (1B), and Smol (500 M) performed worse than random guessing with correct answer percentages lower than 20%. Granite Vision 3.2, InternVL 2.5 (2B) and Smol (2B) performed slightly above random guessing with accuracies of 23.0%, 22.6%, and 25.4%, respectively. The InternVL 2.5 and Qwen 2.5 VL model families exhibited a consistent improvement in accuracy with increased model complexity; for instance, InternVL 2.5 (4B) answered correctly in 32.6% of cases, and InternVL 2.5 (8B) showed slight improvement with an accuracy of 35.0%. The Qwen 2.5 VL (3B) model achieved a 35.6% accuracy, while Qwen 2.5 VL (7B) improved further to 40.4%. The Phi4 Multi (5B) and Gemma 3 (4B) models reached an accuracy of 33.4% and 35.5%, respectively. Notably, none of the open VLMs could compete with GPT-4o, which correctly answered more than two-thirds (68.1%) of challenge questions.

EDs and ICUs, yet none of the evaluated open-source models exhibited the diagnostic accuracy necessary for real-world application in such critical environments.

To better understand the performance relationships and correctness patterns among these VLMs, we computed Phi coefficients to quantify the correlation in their responses (Fig. 4). Strong correlations within model families, such as InternVL and Qwen, demonstrate consistent correctness patterns as model complexity increases. However, models like DeepSeek VL2 Tiny (1B) and Smol (500 M, 2B) have distinctive correctness patterns, likely due to their smaller size and limited capacity.

This performance gap highlights several barriers to the adoption of open-source VLMs for acute care. One notable limitation is the lack of specialized training on medical datasets tailored to ED and ICU settings. The models evaluated in this study were trained on generic multimodal datasets and were not optimized for clinical diagnosis, let alone the unique demands of high-stakes, time-sensitive decision-making in acute care environments. Moreover, the relatively small model sizes of certain open-source VLMs, while enabling deployment on local or mobile hardware, appear to compromise accuracy due to reduced parameter capacity and inferior performance on complex pattern recognition. These limitations suggest that while open-source models could eventually democratize access to advanced diagnostic tools, improved architectures and targeted training strategies are required to close the accuracy gap.

Despite these challenges, the performance of proprietary systems like GPT-4o demonstrates the immense potential of VLMs for acute care. GPT-4o's consistent accuracy across difficulty levels indicates that large-scale models, even in a zero-shot setting without task-specific tuning, can deliver substantial diagnostic value. However, GPT-4o's performance may vary with task design and dataset selection. For instance, Ueda et al.[13] reported up to 98% accuracy on a very selective subset of the NEJM Image Challenge that used only textual descriptions. These differences underscore the importance of contextual factors (e.g., imaging data, question format) and caution against overgeneralizing our findings. In our study, GPT-4o exceeded the overall performance of human responders, indicating considerable potential as decision-support tools in EDs and ICUs. While their diagnostic accuracy may improve patient care efficiency, fully alleviating clinician workload often entails additional features, such as automating documentation and administrative tasks that fall outside this study's scope. Moreover, concerns regarding data privacy, cost, and scalability still limit the widespread adoption of proprietary systems. Consequently, open-source initiatives offer an important avenue for developing ethical and accessible AI solutions tailored to diverse healthcare settings.

If refined for clinical deployment, VLMs could potentially benefit ED and ICU workflows by assisting with triage, aiding in the detection of life-threatening conditions (e.g., pulmonary embolism, severe infections), and supplementing diagnostic accuracy in overburdened healthcare systems. For example, small VLMs running on edge devices could enable bedside image analysis in resource-limited or rural hospitals. Furthermore, ensemble approaches, where multiple models work collaboratively to improve final predictions, could compensate for the weaknesses of smaller individual systems[14].

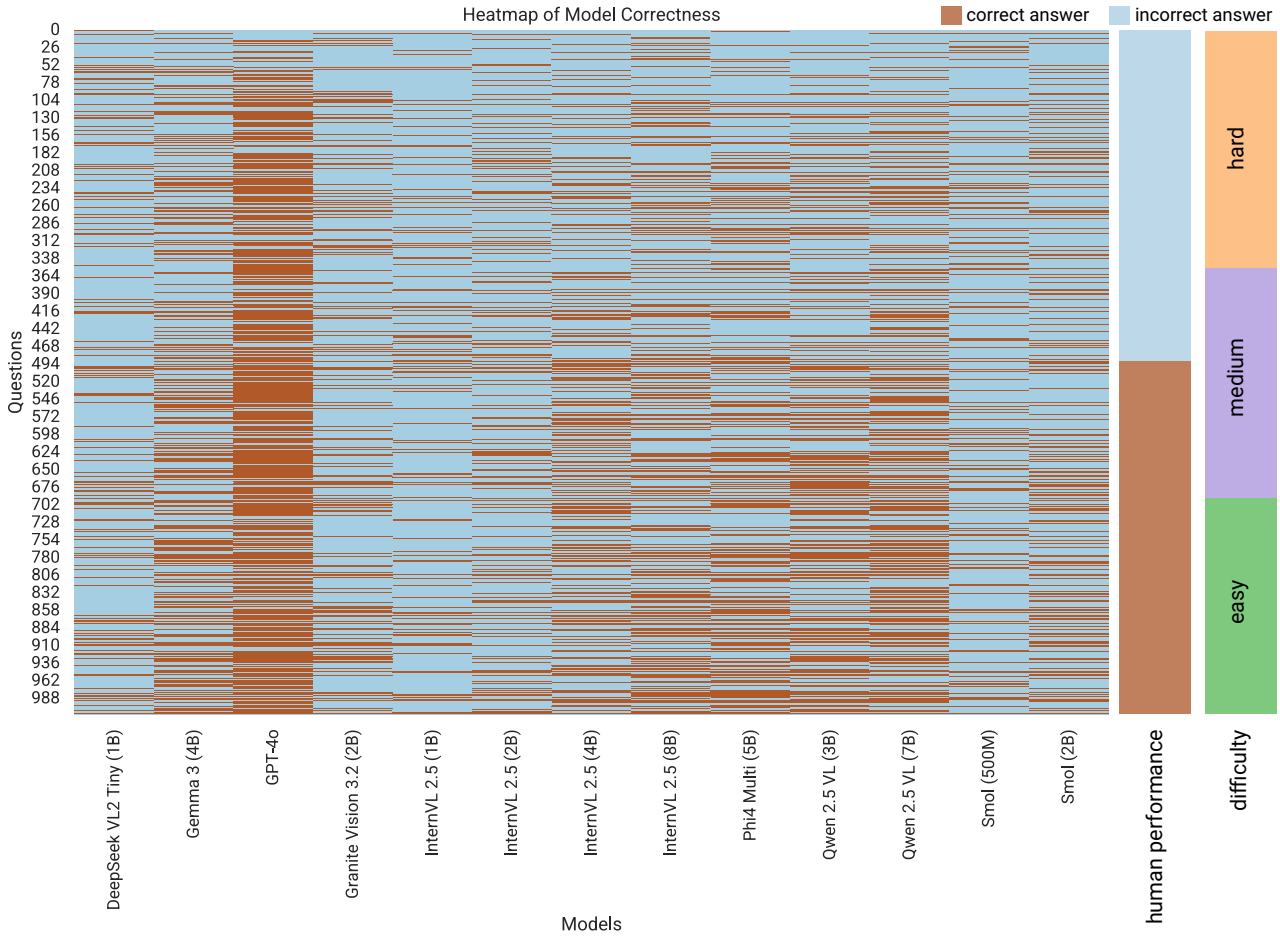

Fig. 3 | Heatmap of model correctness for each challenge question. The colored bars along the right-hand side classify each question's difficulty based on the NEJM human responder accuracy, ranging from 'hard' (orange, ≤44%) at the top, to 'medium' (purple, 45–55%), and 'easy' (green, ≥56%) at the bottom. Each row corresponds to a single question, and each column corresponds to one of the evaluated VLMs. A dark cell indicates that the model selected the correct multiple-choice answer; a blue cell indicates that the model's final answer was incorrect. Humans were categorized as giving the correct answer if more than 50% of NEJM readers answered correctly. The distribution of correct answers across easy, medium, and hard categories remained relatively stable, indicating that the models' capabilities were consistent irrespective of question difficulty as perceived from a human point of view. Despite the accuracy improvements within certain model families, we noted inconsistencies. For example, the smaller InternVL 2.5 (1B) answered some challenges correctly that the larger InternVL 2.5 (2B) did not. Additionally, DeepSeek VL2 Tiny (1B) and InternVL 2.5 (1B) performed comparably and below random guessing, yet their answering patterns showed little overlap.

The correlation analysis showed that models within the same family (e.g., InternVL, Qwen) tended to produce similar responses (Fig. 4), while models with low mutual phi coefficients (e.g., DeepSeek vs. InternVL or GPT-4Q vs. Smol) generated more diverse outputs, suggesting to be good candidates for ensembling, as they likely make different types of predictions, thereby enhancing the overall robustness and accuracy when combined. Another key pathway involves the integration of multimodal data, where VLMs process not only imaging but also patient history, laboratory results, and real-time vitals to offer a more comprehensive and actionable assessment of patient health. However, for these innovations to succeed, rigorous validation and regulatory certification are essential to ensure diagnostic reliability and patient safety.

In real-time applications, particularly within chaotic environments like EDs and ICUs, explainability and interpretability must also be prioritized. Diagnostic output must be easy to understand and actionable for clinicians under pressure. Future efforts should address these practical considerations while also working to mitigate potential biases that may occur in underrepresented populations. To foster equitable AI solutions for all patients, models must be trained on diverse datasets that reflect the population-level variation seen in emergency and acute care settings[15].

The NEJM Image Challenge dataset's inclusion of realistic and diverse cases provides a robust starting point for evaluating VLMs for emergency and critical care applications. However, there are limitations to the current benchmarking approach. While the dataset typically includes clear ground-truth answers, real-world ED/ICU scenarios often involve iterative problem-solving and require clinicians to generate differential diagnoses rather than selecting from multiple-choice options. Future studies should extend these evaluations by incorporating open-ended tasks and additional multimodal benchmarks that better reflect the complex, dynamic nature of clinical reasoning in acute care[16]. It is also important to note that while efforts were made to avoid inadvertent data contamination, the possibility remains that some models may have been trained on publicly accessible medical challenges. Greater transparency in the disclosure of training datasets could address such concerns and bolster the validity of future benchmarking studies[17].

These findings also mirror results reported in non-emergency medical contexts, such as dermatology or radiology screening, where tasks typically involve broader training data and less time-sensitive decision-making. In those domains, smaller open-source VLMs have performed relatively well[18], underscoring the importance of fine-tuning model architectures and training regimens to the complexity and urgency of different clinical scenarios.

In conclusion, this study demonstrates the potential and limitations of small open-source VLMs for emergency and critical care diagnostics. While

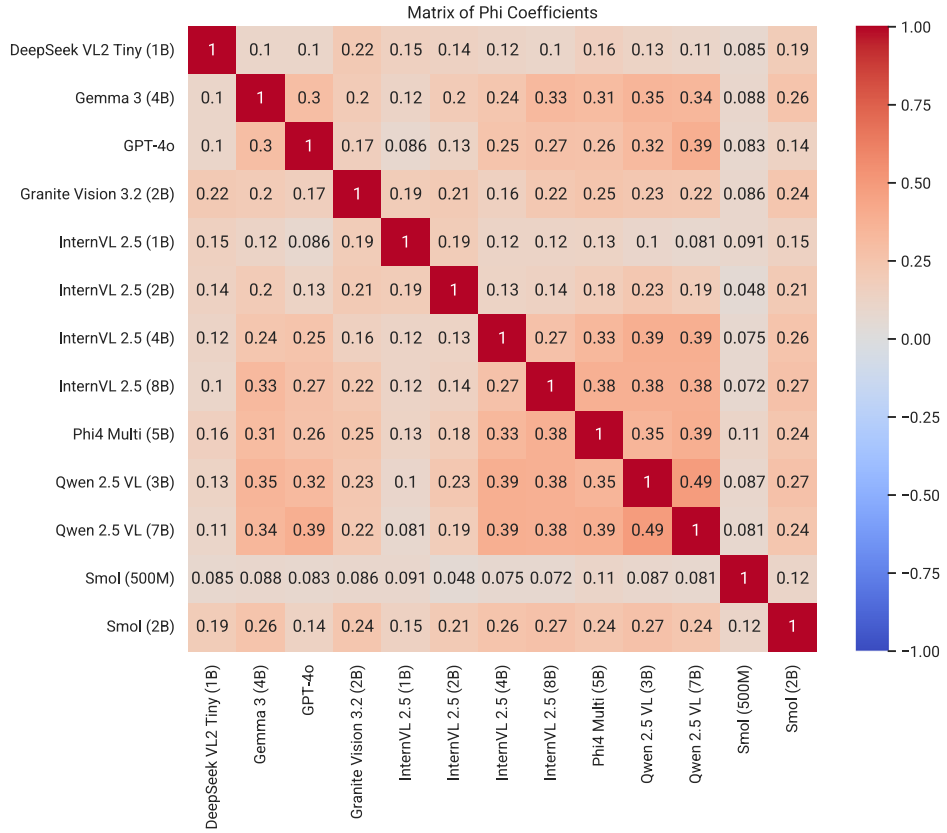

Fig. 4 | Correlation matrix of model response patterns across VLMs. The heatmap displays Phi coefficients (ranging from −1 to 1) quantifying the similarity in correctness patterns among vision-language models, with higher values indicating greater similarity. Strongest correlations (up to 0.5) occur within model families such as InternVL (1B–8B) and Qwen (3B, 7B), reflecting consistent intra-family performance. Smol models also correlate well internally but diverge from others. Granite Vision 3.2 (2B), Phi4 Multi (5B), and Gemma 3 (4B) show moderate correlations with larger InternVL and Qwen models. DeepSeek VL2 Tiny (1B) exhibits low correlation with the similar performing Smol (500 M), suggesting distinctive response patterns.

their modest accuracy renders them insufficient for clinical use at present, refining these models through specialized training, larger architectures, and ensemble approaches holds significant promise for enhancing the reliability and accessibility of AI-driven diagnostic tools. As these technologies mature, their ability to integrate multimodal data streams, reduce diagnostic errors, and support time-sensitive workflows could transform acute care delivery in EDs and ICUs. Continued research and development are essential to bridge the gap between AI innovation and practical implementation in these critical environments.

Methods

VLMs

In this study, we evaluated the ability of several widely used open-source VLMs, including Deepseek VL2 Tiny (1B)[19], Gemma 3 (4B)[20], Granite Vision 2.5 (2B)[21], InternVL 2.5 (1B, 2B, 4B, and 8B)[22], Phi4 Multimodal (5B)[23], Qwen 2.5 VL (3B and 7B)[24], and Smol (500 M and 2.2B)[25]. All these models are less than 6 months old as of April 2025 and were chosen based on their strong performance in benchmarks and popularity on the HuggingFace platform. Additionally, we compared them against GPT-4o[26], a proprietary model, to provide a comprehensive understanding of their capabilities in the field of medical diagnostics.

The values such as 1B, 3B, and 8B represent the number of model parameters (e.g., 1B = 1 Billion), which indicates the model's capacity. Generally, a higher number of parameters can lead to better performance, as the model can capture more intricate patterns in the data. GPT-4o is estimated to have more than 1.8 trillion parameters, although there is no confirmed information on its exact size.

All evaluated VLMs use a two-stage process of extracting visual features from a dedicated vision backbone (e.g., a convolutional or transformer-based encoder) and textual features from a language module, then mapping these features into a shared embedding space via cross-modal attention or contrastive learning[27]. Smaller models (e.g., DeepSeek VL2 Tiny (1B)) generally have fewer parameters and rely on light-weight encoders, which can limit their capacity to learn complex multimodal relationships. In contrast, larger architectures (e.g., Qwen 2.5 VL (8B), GPT-4o) incorporate additional parameters and more advanced bridging mechanisms, enabling more nuanced alignment between image and text features. While exact architectural details are proprietary or unavailable for some models, differences in training data scale, parameter count, and encoder–decoder design can lead to variations in their capability to handle multimodal tasks.

Experiments were conducted on a local machine using a Nvidia A10G graphics card with 24GB RAM, although many of the evaluated models have lower computational requirements. VLMs with 1B or fewer parameters can potentially run on smartphones or edge devices, enabling deployment in resource-constrained environments. In contrast, larger models up to 8B parameters and beyond typically require more dedicated hardware, such as high-performance GPUs, to ensure efficient inference.

We utilized the Huggingface library with Python to load the quantized models and manage text generation parameters. We presented to each model the medical image, relevant background information (if provided), the question and the five possible multiple-choice answers, of which only one was correct, as shown in the NEJM Image Challenge.

To maintain consistency and ensure that each model returned a definitive answer, we prompted the models to generate a single diagnosis using the following instruction (with bracketed text replaced by the case-specific content): "You are a medical expert. Use the image and the description to choose the correct answer. This is the description: {case_description}. Please choose the correct diagnosis among these options: {choice_1}, {choice_2}, {choice_3}, {choice_4}, {choice_5}. Please provide only the answer with no additional explanations. The correct answer is always among the options. You must always return one of the possible choices as an answer."

We relied on the default or recommended (and relatively low-temperature) settings for each VLM that yield near-deterministic outputs and prompted each model to provide only one definitive diagnosis, accuracy metrics in this study reflect a single run for each VLM since multiple runs did not provide different results.

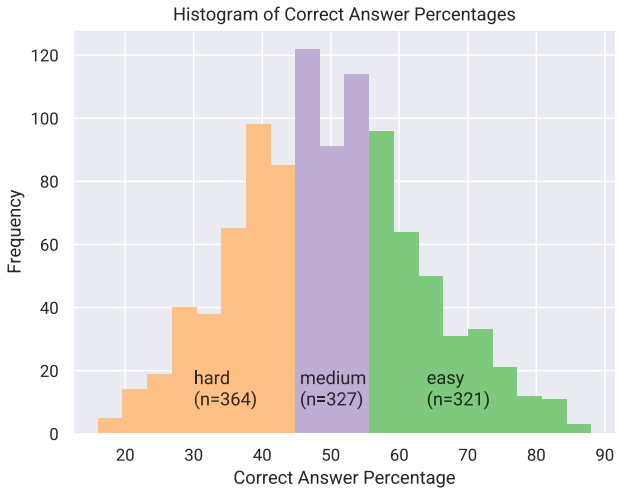

Fig. 5 | Distribution of correct answer percentages for the NEJM Image Challenge by the readers. The histogram shows three difficulty categories: "hard" (orange, 0–44%), "medium" (purple, 45–55%), and "easy" (green, 56–100%) and the number of questions in each category.

NEJM image challenge data

The data utilized in this study were sourced from the NEJM Image Challenge, an educational feature that has been in continuous operation since its inception in 2005. This weekly challenge engages medical professionals and students in diagnostic exercises using real clinical images to enhance their diagnostic skills.

As of March 2025, the NEJM Image Challenge contains a total of 1012 image-text pairs. Each pair consists of an image and its corresponding diagnostic query, along with five multiple-choice options. This dataset offers comprehensive coverage of various medical conditions, scenarios, and educational challenges. We extracted the complete set of images, questions, correct answers, and the proportion of publicly recorded responses for each option for use in this study. The average percentage of reader votes that answered the medical cases correctly was 49.6%. Following Jin et al.[16] we categorized the challenges into three difficulty levels: "easy" for a 56–100% correct answer rate, "medium" for 45–55%, and "hard" for 0–44%, based on the proportion of correct answers from NEJM users. This leads to a relatively even distribution of questions in each category, $n = 321$ in "easy", $n = 327$ of "medium" difficulty, and $n = 364$ "hard" questions (see Fig. 5).

Typically, the diagnostic queries in the challenge involve identifying the correct diagnosis based on the visual cues provided by the clinical images and the accompanying patient information. Occasionally, the questions may focus on correct treatment options or other medical inquiries, broadening the scope of diagnostic evaluation. Sometimes, the questions may be short and generic, such as "What is the most likely diagnosis?" without providing relevant clinical information but always presented with an image. The longest case description consisted of 182 words.

Data availability

The data for this study is publicly available at the NEJM homepage.

Received: 23 April 2025; Accepted: 24 June 2025;
Published online: 10 July 2025

References

1. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).
2. Zhang, Z. & Ni, H. Critical care studies using large language models based on electronic healthcare records: a technical note. *J. Intens. Med.* https://doi.org/10.1016/j.jointm.2024.09.002 (2024).
3. Koga, S. & Du, W. From text to image: challenges in integrating vision into ChatGPT for medical image interpretation. *Neural Regen. Res.* **20**, https://journals.lww.com/nrronline/fulltext/2025/02000/from_text_to_image__challenges_in_integrating.25.aspx (2025).
4. Clusmann J. et al. The future landscape of large language models in medicine. *Commun. Med.* **3**, https://doi.org/10.1038/s43856-023-00370-1 (2023).
5. Bedi, S. et al. Testing and evaluation of health care applications of large language models. *JAMA* **333**, 319 (2025).
6. Sonoda, Y. et al. Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases. *Jpn J. Radiol.* https://doi.org/10.1007/s11604-024-01619-y (2024).
7. Kim, S. H. et al. Benchmarking the diagnostic performance of open source LLMs in 1933 Eurorad case reports. *NPJ Digit. Med.* **8**, https://www.nature.com/articles/s41746-025-01488-3 (2025).
8. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180, https://doi.org/10.1038/s41586-023-06291-2 (2023).
9. Liu, X. et al. A generalist medical language model for disease diagnosis assistance. *Nat. Med.* http://www.ncbi.nlm.nih.gov/pubmed/39779927 (2025).
10. Chua, M. T. et al. The role of artificial intelligence in sepsis in the Emergency Department: a narrative review. *Ann. Transl. Med.* **13**, 4–4 (2025).
11. Dennstädt, F., Hastings, J., Putora, P. M., Schmerder, M. & Cihoric, N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *NPJ Digit Med* **8**, 143 (2025).
12. Riedemann, L., Labonne, M. & Gilbert, S. The path forward for large language models in medicine is open. *NPJ Digit. Med.* **7**, https://doi.org/10.1038/s41746-024-01344-w (2024).
13. Ueda, D. et al. Evaluating GPT-4-based ChatGPT's clinical potential on the NEJM quiz. *BMC Digital Health* **2**, https://bmcdigitalhealth.biomedcentral.com/articles/10.1186/s44247-023-00058-5. (2024).
14. Yang, H. et al. One LLM is not enough: harnessing the power of ensemble learning for medical question answering. *J. Med. Internet Res*. https://doi.org/10.2196/70080 (2024).
15. Hager, P. et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat. Med.* https://doi.org/10.1038/s41591-024-03097-1 (2024).
16. Jin, Q. et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine. *NPJ Digit. Med.* **7**, https://www.nature.com/articles/s41746-024-01185-7 (2024).
17. Longpre, S. et al. Data authenticity, consent, and provenance for AI are all broken: what will it take to fix them? An MIT exploration of generative AI. https://doi.org/10.21428/e4baedd9.a650f77d (2024).
18. Tanno, R. et al. Collaboration between clinicians and vision–language models in radiology report generation. *Nat. Med* https://doi.org/10.1038/s41591-024-03302-1 (2024).
19. Wu, Z. et al. DeepSeek-VL2: mixture-of-experts vision-language models for advanced multimodal understanding. Preprint at http://arxiv.org/abs/2412.10302 (2024).
20. Gemma Team. Gemma 3 Technical Report. https://goo.gle/Gemma3Report (2025).
21. Granite Vision Team, Karlinsky, L. et al. Granite Vision: a lightweight, open-source multimodal model for enterprise Intelligence. Preprint at http://arxiv.org/abs/2502.09927 (2025).
22. Chen, Z. et al. InternVL: scaling up vision foundation models and aligning for generic visual-linguistic tasks. https://github.com/OpenGVLab/InternVL (2024).
23. Abdin, M. et al. Phi-4 technical report. Preprint at http://arxiv.org/abs/2412.08905 (2024).
24. Wang, P. et al. Qwen2-VL: enhancing vision-language model's perception of the world at any resolution. Preprint at http://arxiv.org/abs/2409.12191 (2024).
25. Marafioti, A. et al. Smolvlm: redefining small and efficient multimodal models. arXiv preprint arXiv:2504.05299 (2025).
26. Hurst, A. et al. GPT-4o system card. Preprint at http://arxiv.org/abs/2410.21276 (2024).

27. Bordes, F. et al. An introduction to vision-language modeling. Preprint at http://arxiv.org/abs/2405.17247 (2024).

Acknowledgements
This study was sponsored by Novartis Pharma.

Author contributions
C.F.K. conceptualized and designed the study, collected and analyzed the data and wrote the initial draft of the manuscript. T.M., B.M.E., J.N.K., and B.G. contributed to the design of the experiments, provided critical review, and contributed to the manuscript.

Competing interests
C.F.K. and B.G. report employment and stock ownership from Novartis Pharma. B.G. is a Strategic Advisory Board Member at Fraunhofer IZI-BB. B.G. serves as an Associate Editor of npj Digital Medicine and was not involved in the review or decision-making process for this manuscript. J.N.K. declares consulting services for Bioptimus, France; Panakeia, UK; AstraZeneca, UK; and MultiplexDx, Slovakia. Furthermore, J.N.K. holds shares in StratifAI, Germany, Synagen, Germany, Ignition Lab, Germany; J.N.K. has received an institutional research grant by GSK; and has received honoraria by AstraZeneca, Bayer, Daiichi Sankyo, Eisai, Janssen, Merck, MSD, BMS, Roche, Pfizer, and Fresenius. T.M. and B.M.E. have nothing to declare.

Additional information
Correspondence and requests for materials should be addressed to Christoph F. Kurz.