Can Agents Distinguish Visually Hard-to-Separate Diseases in a Zero-Shot Setting? A Pilot Study

Zihao Zhao[0009−0001−8044−7683], Frederik Hauke[0000−0003−3434−5720], Juliana De Castilhos[0000−0001−9966−8657], Sven Nebelung[0000−0002−5267−9962], and Daniel Truhn (✉)[0000−0002−9605−0728]

Department of Diagnostic and Interventional Radiology, University Hospital Aachen, 52074 Aachen, Germany
dtruhn@ukaachen.de

Abstract. The rapid progress of multimodal large language models (MLLMs) has led to increasing interest in agent-based systems. While most prior work in medical imaging concentrates on automating routine clinical workflows, we study an underexplored yet clinically significant setting: distinguishing visually hard-to-separate diseases in a zero-shot setting. We benchmark representative agents on two imaging-only proxy diagnostic tasks, (1) melanoma vs. atypical nevus and (2) pulmonary edema vs. pneumonia, where visual features are highly confounded despite substantial differences in clinical management. We introduce a multi-agent framework based on contrastive adjudication. Experimental results show improved diagnostic performance (an 11-percentage-point gain in accuracy on dermoscopy data) and reduced unsupported claims on qualitative samples, although overall performance remains insufficient for clinical deployment. We acknowledge the inherent uncertainty in human annotations, conservative binary setting, and the absence of clinical context, which further limit the translation to real-world settings. Within this controlled setting, this pilot study provides preliminary insights into zero-shot agent performance in visually confounded scenarios. Our code is available at https://github.com/TruhnLab/Contrastive-Agent-Reasoning.

Keywords: Visually confounded diseases · Multi-agent system · Multimodal Large Language Model.

1 Introduction

Agents are systems that receive observations and perform goal-directed actions (e.g., making decisions or plans) to optimize specific objectives [1]. In recent years, multimodal large language models (MLLMs) [3,21,19,20,7] have advanced rapidly in capability, making them well-suited to serve as such agents. Hence, a growing number of studies [24] have adopted MLLMs as agents and integrated them into clinical practice, aiming to automate routine workflows or assist physicians in decision-making. However, most existing systems [24] assume always-on

human-agent collaboration, while some recent studies [23,18] have questioned this assumption and suggested selective collaboration.

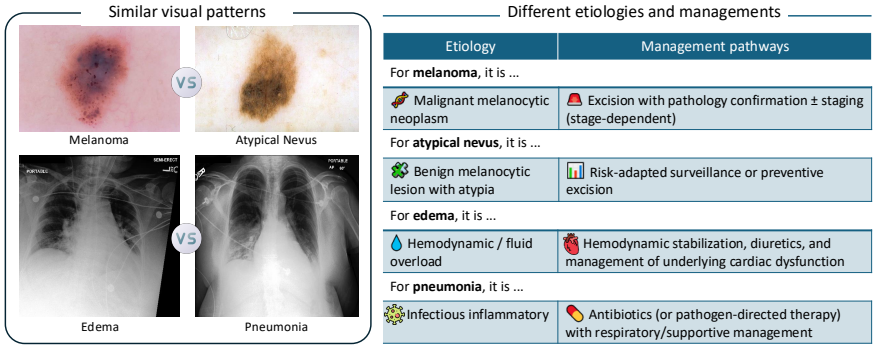

Fig. 1. Illustration of hard-to-separate disease pairs. Despite highly overlapping visual patterns, their typical etiologies and managements differ significantly, which makes imaging-only differentiation challenging and high-stakes.

Motivated by this, we focus on a less explored yet clinically critical setting: distinguishing visually hard-to-separate diseases with distinct etiologies and management pathways. Fig. 1 provides a straightforward illustration. In dermoscopy, distinguishing melanoma from atypical nevus is a well-recognized diagnostic challenge [10] because both are melanocytic lesions and can share asymmetry and irregular borders. As a fatal cancer, concern about melanoma is a common reason for dermatology consultation and skin checks [15]. Likewise, on chest radiographs, edema and pneumonia are frequently confused because both present with lung opacities and diffuse haziness [6]. This makes separation challenging for inexperienced radiologists. Despite visual similarity, their management methods are different. This discrepancy naturally raises a critical question: can current MLLM-based agents distinguish such visually confounded diseases in a zero-shot manner, and thus assist junior clinicians?

However, current MLLM-based agents are still far from satisfactory. Under high-ambiguity scenarios, a single agent could prematurely favor one hypothesis, and produce overconfident statements to support this hypothesis [8]. Recent efforts to alleviate this issue rely on additional annotated data for fine-tuning [13,14], repeated sampling to assess response consistency [25,27], agent collaboration with tool-use mechanisms [5], or training recipes designed to mitigate overconfidence [9]. Such approaches, however, either introduce external tool dependencies or require additional training, and are not well-aligned with our zero-shot setting. To this end, we propose a novel multi-agent system named **C**ontrastive **A**gent **RE**asoning (CARE). The underlying philosophy is straightforward: even experienced human experts reason by contrast [2]. Specifically, explaining why the case supports one side and why it opposes the other. In

CARE, one agent argues for melanoma while another argues for atypical nevus, and a third agent acts as a judge and summarizes their contrastive arguments to make the final diagnosis.

To summarize, our main contributions are listed as follows:

- To the best of our knowledge, this is among the first studies to benchmark MLLM-based agents on visually confounded diseases in a zero-shot setting.
- We propose CARE, a novel multi-agent system that improves agent performance by explicitly structuring disagreement, without additional training.
- Extensive benchmarking across two imaging modalities demonstrates statistically significant improvements over baselines ($p < 0.0001$ for dermoscopy; $p < 0.001$ for chest X-rays, McNemar and permutation tests), yet overall performance remains below the requirements for clinical use.

2 Method

We first formulate the problem setting in Section 2.1. Then, we detail the principle of CARE (Section 2.2). A brief analysis is provided in Section 2.3.

2.1 Problem Formulation

In this study, we focus on image-only prediction of report-derived labels under a forced exclusive-or (XOR) setting. Let $x \in \mathcal{X}$ denote a medical image (e.g., dermoscopy or chest radiograph), and let $y \in \{A, B\}$ denote the corresponding diagnostic label extracted from a human-written report. The XOR criterion enforces mutual exclusivity between the two labels, thereby formulating the task as a controlled binary classification problem. Although this setup simplifies evaluation, it does not reflect the potential co-occurrence of conditions (e.g., edema and pneumonia) in real clinical practice. Our setting is intentionally constrained to zero-shot inference: no task-specific fine-tuning and no annotated supervision. The goal is to output a diagnostic decision, together with visual evidences and an interpretable reasoning path. This setting differs from standard image classification, as a single probability score only provides limited decision support.

2.2 Contrastive Multi-Agent Reasoning

As illustrated in Fig. 2, CARE is training-free and does not require external tools or fine-tuning, but uses a structured multi-call inference with three roles. Two disease-specific agents independently generate evidence conditioned on opposing diagnostic hypothesis, while a third agent adjudicates between their arguments.

Role-conditioned evidence generation. Each disease-specific agent operates under a strict role condition. It must interpret the image solely from the perspective of its assigned diagnosis and enumerate visual evidence supporting that hypothesis. Agents are prohibited from making a final diagnosis. This role conditioning ensures that each agent is dedicated solely to evidence generation, rather than decision making. As a result of this design choice, unsupported

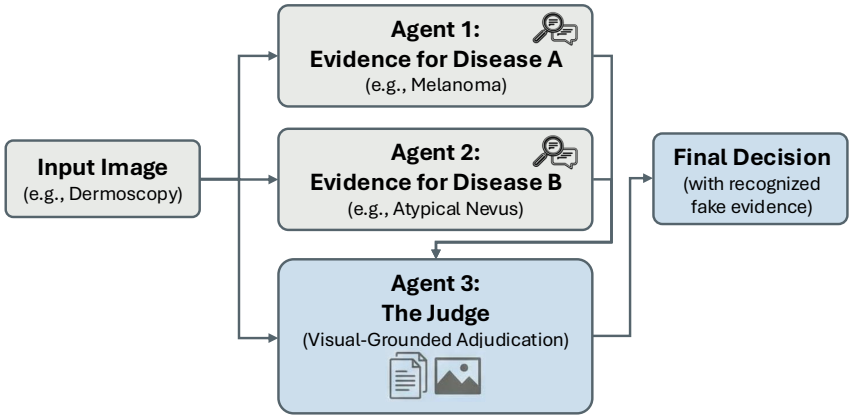

Fig. 2. The overview of **Contrastive Agent Reasoning (CARE).** Two disease-specific agents generate opposing evidence from the same input image (e.g., melanoma vs. atypical nevus). A judge agent adjudicates the arguments, flags unsupported evidence, and outputs the final diagnosis in a training-free, zero-shot setting.

evidence becomes hypothesis-consistent but potentially image-inconsistent. Evidence generated for disease A could contradict or undermine disease B, and vice versa. These inconsistencies enable contrastive adjudication.

Visual-grounded judgment. The judge agent receives three inputs: the original medical image x, the evidence set E_A generated under hypothesis $y = A$, and the evidence set E_B generated under hypothesis $y = B$. Its task is to evaluate the plausibility of each evidence set by grounding the claims in the image. The judge agent performs three functions: (i) cross-checking evidence against the image, (ii) identifying unsupported or contradictory claims, and (iii) weighing the remaining contrastive arguments to reach a final diagnosis. Importantly, the judge does not introduce new medical evidence. It only evaluates existing claims.

2.3 Why Contrastive Reasoning Benefits

The rationale of contrastive reasoning can be illustrated probabilistically. A single agent implicitly commits to a decision

$$\hat{y} = \arg \max_{y \in \{A,B\}} p(y \mid x),$$

which becomes unstable when $p(A \mid x) \approx p(B \mid x)$, as is common in visually confounded cases. Hallucinated evidence, while often self-consistent within a single hypothesis, tends to become inconsistent across mutually exclusive hypotheses. Hence, instead of committing to a single hypothesis, CARE explicitly generates hypothesis-conditioned explanations:

$$E_A \leftarrow \text{Agent}_1(x), \qquad E_B \leftarrow \text{Agent}_2(x).$$

A judge agent then evaluates the visual consistency of these competing explanations. Let $\mathcal{S}(x, E)$ denote an implicit visual-consistency assessment performed by the judge agent. Conceptually, the final decision can be expressed as

$$\hat{y} = \arg \max_{y \in \{A,B\}} \Big(\mathcal{S}(x, E_y) - \mathcal{S}(x, E_{\neg y}) \Big).$$

Rather than relying on repetitive self-check [25] or stochastic sampling [27], CARE induces contrast through structured role assignment. By forcing explicit contrast between competing explanations, CARE distributes reasoning across competing hypotheses, mitigating premature commitment under visual ambiguity. Note that CARE is implemented via structured prompting only. It does not compute explicit likelihoods or numerical functions in practice.

3 Experimental Results

3.1 Dataset Curation and Implementations

In this section, we curate two binary classification datasets from existing public datasets based on the following pipeline.

For **Melanoma vs. Atypical Nevus**, we curate dermoscopy images from the derm7pt dataset [12]. We first enforce an XOR criterion, retaining only studies labeled as either atypical nevus or melanoma. We then exclude congenital and recurrent nevi, as these subtypes are often very difficult without clinical context. To mitigate class imbalance, we subsample atypical nevi, resulting in a final set of 509 studies with 257 atypical nevi and 252 melanomas. For **Edema vs. Pneumonia**, we curate chest X-rays from the MIMIC-CXR dataset [11]. After the XOR criterion, we obtain 2,907 candidate studies with exclusive edema or pneumonia labels. Since report-derived labels can be noisy, we further exclude studies with low-confidence diagnostic statements by identifying expressions like "can not be excluded" based on associated radiology reports, yielding a final set of 1,739 studies (878 edema and 861 pneumonia).

Implementations. All curated data are used exclusively for evaluation (test-only) in this study. We provide no external tools for MLLM-based agents to fairly compare their capabilities. Experiments with open-source MLLMs are conducted on a single NVIDIA H100 GPU (96 GB). Due to data privacy requirements, Gemini, instead of GPT, is selected as the default closed-source MLLM in our study. We obtain API access to Gemini models via the Vertex AI platform, as Physionet requested. Accuracy (ACC), F1-score (F1), and Youden-Index (Youden) [26], defined as Sensitivity + Specificity − 1, are reported. CARE is built upon Gemini-3-Flash by default in this study. For Gemini-3-Flash vs. CARE and Gemini-3-Pro vs. CARE, statistical significance is evaluated, using McNemar [16] (for ACC) and permutation tests [4] (for F1 and Youden).

3.2 Quantitative Analysis

Table 1 presents a comprehensive comparison across three categories of models: CLIP-based vision-language models, open-source, and closed-source MLLMs.

Table 1. Benchmarking results on visually confounded diagnostic tasks. Gray : best single-agent method. **Bold**: best performance across all methods.

Method	Melanoma vs. Atypical Nevus			Edema vs. Pneumonia		
	ACC	F1	Youden	ACC	F1	Youden
SigLIP2 [22]	0.611	0.606	0.219	0.502	0.349	0.006
BiomedCLIP [28]	0.607	0.569	0.209	0.557	0.554	0.114
MedVLM-R1 [17]	0.505	0.368	0.001	0.498	0.336	-0.003
InternVL3-14B [3]	0.549	0.534	0.071	0.502	0.361	0.002
Gemma-3-12B [19]	0.499	0.427	0.091	0.476	0.455	-0.041
Qwen3-VL-8B [21]	0.661	0.656	0.365	0.559	0.542	0.119
Qwen3-VL-32B [21]	0.698	0.670	0.347	0.564	0.504	0.116
GLM-4.6V-Flash [20]	0.718	0.714	0.426	0.552	0.513	0.088
Gemini-3-Flash [7]						
Baseline	0.665	0.630	0.328	0.602	0.539	0.197
Self-Check† (2×)	0.701	0.681	0.400	0.606	0.546	0.200
Self-Check† (3×)	0.719	0.704	0.436	0.610	0.551	0.208
Majority-Vote‡ (3×)	0.686	0.660	0.376	0.601	0.533	0.195
Blind-CARE*	0.739	0.729	0.476	0.618	0.563	0.228
CARE	**0.776**	**0.769**	**0.552**	0.646	0.619	0.287
Gemini-3-Pro [7]	0.741	0.734	0.482	**0.709**	**0.709**	**0.419**

† Multi-pass self-revision; ‡ Independent sampling with majority aggregation; * Judge agent has no access to the original image.

Overall performance reveals substantial challenges. CLIP-based models (SigLIP2 and BiomedCLIP) achieve modest performance on both tasks, indicating that simple vision-language alignment without reasoning capabilities is insufficient for these visually confounded diagnostic scenarios. Other single-agent systems show better performance but remain largely unsatisfactory, with most of them achieving only 50-70% accuracy. This highlights the difficulty of distinguishing visually confounded diseases.

CARE yields consistent improvements. Built upon Gemini-3-Flash, CARE framework demonstrates notable performance gains. On the melanoma vs. atypical nevus task, CARE achieves 77.6% accuracy with a Youden Index of 0.552, representing an improvement of over 11 percentage points compared to the single-agent Gemini-3-Flash baseline (66.5% accuracy, 0.328 Youden Index). Notably, the difference between CARE and Gemini-3-Pro is not statistically significant on this dataset, based on McNemar (p-value = 0.192) and permutation (p-value = 0.192 for both F1 and Youden) tests. On the edema vs. pneumonia task, while CARE still outperforms its base model (64.6% vs. 60.2%), it falls short of Gemini-3-Pro's 70.9 %. Nevertheless, the improvement over Gemini-3-Flash is still significant, with p-value < 0.001 on all three metrics.

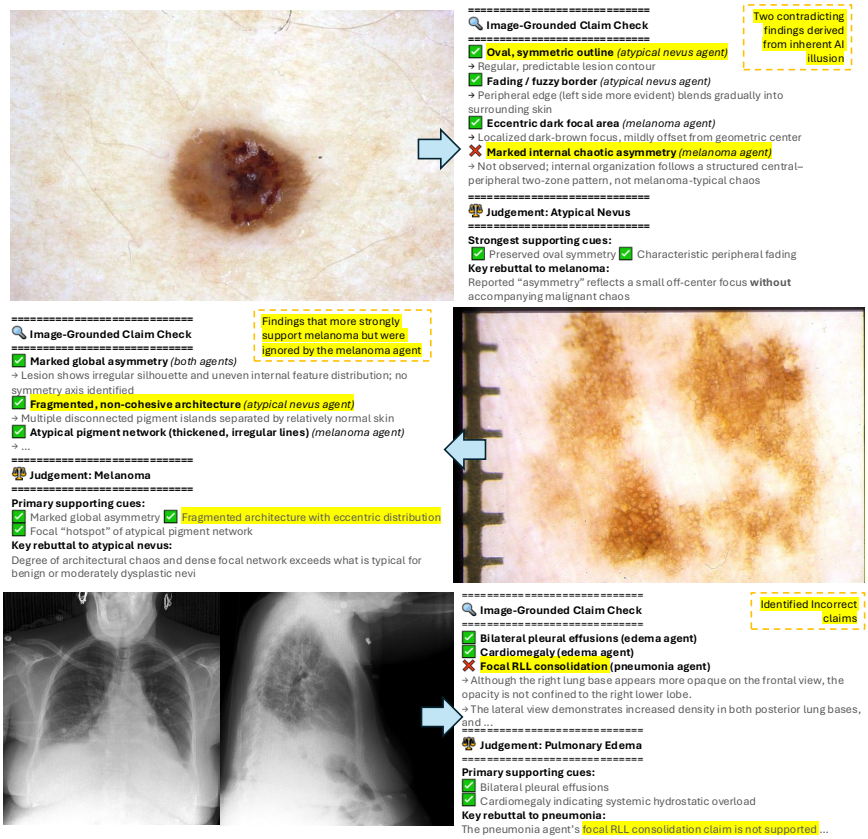

Fig. 3. Representative cases showing how CARE exposes contradictory findings, implements cross-agent evidence recalibration, and verifies claims against the image.

Ablation studies validate the design of CARE. We design three types of ablation variants to justify the contribution of our proposed CARE. Self-Check (2×) and Self-Check (3×) prompt Gemini-3-Flash to perform multi-pass self-reflection before producing a final decision, with two and three sequential reasoning passes, respectively. Majority-Vote (3×) independently queries Gemini-3-Flash three times and aggregates the final prediction via majority voting. These two types of variants show only limited improvement over the single-call baseline, particularly on chest radiographs. Notably, Self-Check (3×) and Majority-Vote (3×) are compute-matched to CARE, as each method invokes the API three times per case. Their marginal gains indicate that performance improvement is not merely attributable to increased sampling or ensembling, but rather to the structured contrastive reasoning mechanism introduced by CARE. Interestingly, Majority-Vote tends to perform worse than Self-Check. One possible explanation is that errors in visually confounded cases may be correlated rather than purely

random. If similar interpretations are repeatedly produced, simple aggregation may provide a limited corrective effect. We further evaluate the third type of variant, Blind-CARE, in which the judge agent has access only to the textual arguments from the two specialist agents. Blind-CARE achieves 73.9% accuracy on melanoma, outperforming the two Self-Check variants but remaining inferior to CARE. This suggests that direct access to visual evidence is essential for detecting fabricated or misinterpreted claims and for effective adjudication.

3.3 Qualitative Analysis

This section illustrates how CARE behaves in representative cases via Fig. 3. First, CARE can detect contradictory findings. In the first case, the melanoma agent claims "marked internal chaotic asymmetry" that conflicts with the lesion's overall symmetric architecture, which the judge explicitly flags and rejects via image-grounded claim checking, yielding the correct atypical nevus decision. Second, CARE enables **cross-agent evidence recalibration**. In visually confounded cases, certain findings may be compatible with both hypotheses. However, their specific morphological characteristics and spatial distribution determine which diagnosis they more strongly support. In our example, a fragmented and non-cohesive architecture is mistakenly interpreted during evidence generation as favoring atypical nevus. The judge re-evaluates this finding and correctly recalibrates its diagnostic weight toward melanoma. Finally, CARE can identify an unsupported localized claim from the pneumonia agent by cross-checking against multi-view evidence, preventing a false pneumonia prediction and correctly concluding pulmonary edema. Overall, CARE reduces ungrounded claims in the presented examples.

4 Discussion and Conclusion

In this study, we benchmark multiple MLLM-based agents and find their performance limited. Some methods even achieve a negative Youden index, indicating performance below chance level. This study has several limitations. First, label quality is imperfect. While part of the dermoscopy data was histologically verified [12], the rest relied on expert diagnosis. Chest X-ray labels were automatically extracted from radiology reports and therefore reflect report impressions rather than an independent reference standard (e.g., CT or clinical adjudication), introducing evaluation uncertainty. Our mutually exclusive setting also conflicts with the fact that patients may have both edema and pneumonia. Second, agents were evaluated without external tools. Future work could investigate whether segmentation models or image retrieval could help improve performance.

In conclusion, we investigate a clinically important yet underexplored topic through imaging-only proxy tasks. We propose CARE, a multi-agent system improving zero-shot diagnostic performance in a training-free setting. Findings suggest that structuring disagreement and image-based verification are essential for performance gain, offering insights for future multi-agent system design.

Nevertheless, our overall benchmarking results underscore that current MLLM-based agents are not yet suitable for clinical translation, highlighting the need for further methodological advances and more rigorous evaluation.

Acknowledgments. This research project is funded by the European Union Research and Innovation Programme (Horizon Europe - European Research Council, SAGMA – GA 101222556). The authors gratefully acknowledge the computing time provided to them at the NHR Center NHR4CES at RWTH Aachen University (project number p0021834). This is funded by the Federal Ministry of Research, Technology and Space , and the state governments participating on the basis of the resolutions of the GWK for national high performance computing at universities (www.nhr-verein.de/unsere-partner). The data used in this publication was managed using the research data management platform Coscine (http://doi.org/10.17616/R31NJNJZ) with storage space of the Research Data Storage (RDS) (DFG: INST222/1261-1) and DataStorage.nrw (DFG: INST222/1530-1) granted by the DFG and Ministry of Culture and Science of the State of North Rhine-Westphalia.

Disclosure of Interests. Daniel Truhn holds shares in StratifAI and Synagen. He has received honoraria from Bayer, AstraZeneca, Philips, Roche, Pfizer, and Gilead. All other authors declare no conflicts of interest.

References

1. Akerkar, R.: Ai agents: From perception to action. In: Artificial Intelligence: Transcending Traditional Paradigms, pp. 303–355. Springer (2026)
2. Buçinca, Z., Swaroop, S., Paluch, A.E., Doshi-Velez, F., Gajos, K.Z.: Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. pp. 1–25 (2025)
3. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
4. Collingridge, D.S.: A primer on quantitized data analysis and permutation testing. Journal of mixed methods research **7**(1), 81–97 (2013)
5. Feng, Y., Du, J., Hong, Y., Wang, Q., Yu, L.: Pass: Probabilistic agentic supernet sampling for interpretable and adaptive chest x-ray reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 30717–30725 (2026)
6. Gluecker, T., Capasso, P., Schnyder, P., Gudinchet, F., Schaller, M.D., Revelly, J.P., Chiolero, R., Vock, P., Wicky, S.: Clinical and radiologic features of pulmonary edema. Radiographics **19**(6), 1507–1531 (1999)
7. Google DeepMind: Gemini 3 (2025), `https://deepmind.google/technologies/gemini/`
8. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 1–55 (2025)

9. Jiang, S., Chen, Y., Song, S., Zhang, Y., Jin, Y., Feng, Y., Wu, J., Liu, Z.: Knowing or guessing? robust medical visual question answering via joint consistency and contrastive learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 325–335. Springer (2025)
10. Jitian, C.R., Frățilă, S., Rotaru, M.: Clinical-dermoscopic similarities between atypical nevi and early stage melanoma. Experimental and Therapeutic Medicine **22**(2), 854 (2021)
11. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
12. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE journal of biomedical and health informatics **23**(2), 538–546 (2018)
13. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13872–13882 (2024)
14. Liang, X., Liu, C., Ma, Z., Wang, D., Jing, B., Wang, Q., Shi, Y.: Anatomical region-guided contrastive decoding: A plug-and-play strategy for mitigating hallucinations in medical vlms. arXiv preprint arXiv:2512.17189 (2025)
15. Mahama, A.N., Haller, C.N., Labrada, J., Idiong, C.I., Haynes, A.B., Jacobs, E.A., Tsevat, J., Pignone, M.P., Adamson, A.S.: Lived experiences and fear of cancer recurrence among survivors of localized cutaneous melanoma. JAMA dermatology **160**(5) (2024)
16. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947)
17. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.: Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 337–347. Springer (2025)
18. Rosbach, E., Ganz, J., Ammeling, J., Riener, A., Aubreville, M.: Automation bias in ai-assisted medical decision-making under time pressure in computational pathology. In: BVM Workshop. pp. 129–134. Springer (2025)
19. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
20. Team, G.V.: Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning (2025), `https://arxiv.org/abs/2507.01006`
21. Team, Q.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
22. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
23. Vaccaro, M., Almaatouq, A., Malone, T.: When combinations of humans and ai are useful: A systematic review and meta-analysis. Nature Human Behaviour **8**(12), 2293–2303 (2024)

24. Wang, W., Ma, Z., Wang, Z., Wu, C., Ji, J., Chen, W., Li, X., Yuan, Y.: A survey of llm-based agents in medicine: How far are we from baymax? arXiv preprint arXiv:2502.11211 (2025)
25. Wienholt, P., Caselitz, S., Siepmann, R., Bruners, P., Bressem, K., Kuhl, C., Kather, J.N., Nebelung, S., Truhn, D.: Hallucination filtering in radiology vision-language models using discrete semantic entropy. arXiv preprint arXiv:2510.09256 (2025)
26. Youden, W.J.: Index for rating diagnostic tests. Cancer **3**(1), 32–35 (1950)
27. Zhang, S., Sambara, S., Banerjee, O., Acosta, J.N., Fahrner, L.J., Rajpurkar, P.: Radflag: A black-box hallucination detection method for medical vision language models. In: Machine Learning for Health (ML4H). pp. 1087–1103. PMLR (2025)
28. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. Nejm Ai **2**(1), AIoa2400640 (2025)