Which Tool Response Should I Trust? Towards Tool-Expertise-Aware Chest X-ray Agent with Multimodal Agentic Reinforcement Learning

Zheang Huai, Honglong Yang, and Xiaomeng Li

Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
zhuaiaa@connect.ust.hk

Abstract. AI agents with tool-use capabilities show promise for integrating the domain expertise of various tools. In the medical field, however, tools are usually AI models that are inherently error-prone and can produce contradictory responses. Existing research on medical agents lacks sufficient understanding of the tools' realistic reliability and thus cannot effectively resolve tool conflicts. To address this gap, this paper introduces a framework that enables an agent to interact with tools and empirically learn their practical trustworthiness across different types of multimodal queries via agentic reinforcement learning. As a concrete instantiation, we focus on chest X-ray analysis and present a tool-expertise-aware chest X-ray agent (TEA-CXA). When tool outputs disagree, the agent experimentally accepts or rejects multimodal tool results, receives rewards, and learns which tool to trust for each query type. Importantly, TEA-CXA extends existing codebases for reinforcement learning with multi-turn tool-calling that focus on textual inputs, to support multimodal contexts effectively. In addition, we enhance the codebase for medical use scenarios by supporting multiple tool calls in one turn, parallel tool inference, and multi-image accommodation within a single user query. Our code framework is applicable to general medical research on multi-turn tool-calling reinforcement learning in multimodal settings. Experiments show that TEA-CXA outperforms the state-of-the-art methods and a comprehensive set of strong baselines. Code is available at https://github.com/zheangh/TEA-CXA.

Keywords: Agent · Tool-expertise awareness · Multimodal agentic reinforcement learning · Chest X-ray.

1 Introduction

Large language models (LLMs) and multi-modal large language models (MLLMs) have demonstrated remarkable capabilities in decision-making in complex interactive environments [17]. Concurrently, task-specific models have shown success in automating various aspects of medical image interpretation, such as Chest X-ray (CXR) report generation [25], classification [5], and phrase grounding [2].

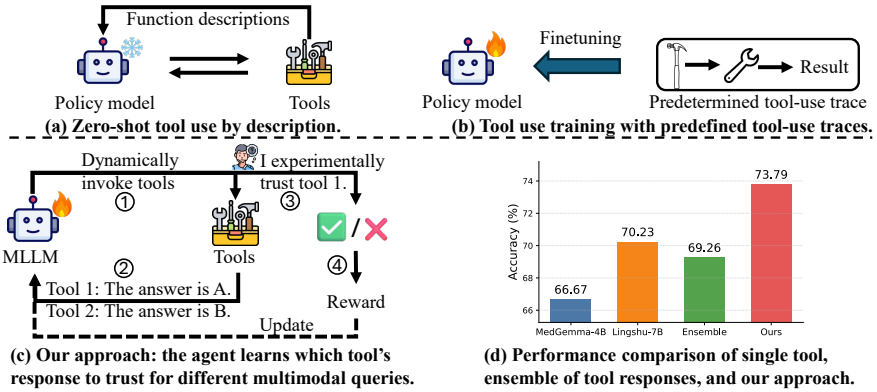

Fig. 1. (a)(b) Previous works use tools in a zero-shot manner, or fine-tune the policy model with pre-specified tool-use traces to enable tool-calling behavior without addressing the realistic reliability of tool outputs. (c) Our approach enables the agent to learn tools' empirical trustworthiness across different queries through multimodal agentic reinforcement learning. (d) Our method outperforms any individual tool and agent-based ensembling of tool outputs.

Medical AI agents with tool-use capabilities have emerged to integrate the domain expertise of various AI models. An agent is driven by a policy model, typically an LLM or MLLM. It can understand user queries, make decisions, and invoke appropriate specialized models, hereafter referred to as "tools", to obtain expert opinions or carry out specific tasks [16, 27, 8].

Current research on medical agents with tool use primarily falls into two categories, as illustrated in Fig. 1(a,b). The first category equips a frozen policy model with a suite of external tools and enables it to invoke these tools and integrate their outputs in a zero-shot manner based solely on their functional descriptions. For example, [6, 14, 11, 29, 7] curate domain-specific expert tools tailored to particular tasks and design agents with a frozen policy model. The second category focuses on fine-tuning the policy model using pre-constructed tool-use traces to enable appropriate tool calling, e.g., functionality-based tool selection and tool argument generation. For instance, [16, 12, 8, 22, 9] construct datasets with complete tool-use traces and fine-tune the policy model to enable it to invoke functionally relevant tools with arguments of correct formats.

However, prior research on medical agents lacks sufficient understanding of the tools' real-world reliability. The tools commonly used in medical settings are AI models that are inherently prone to inaccuracies [1], and may yield contradictory outputs. Their performance also varies across datasets [24]. Relying solely on tool descriptions or pre-constructed tool-use traces prevents agents from having access to tool outputs' practical trustworthiness on the target dataset, leaving them unable to determine which output is correct when conflicts arise. This underscores the need for medical agents to develop a nuanced understanding of each tool's real-world strengths and limitations across different types of

multimodal queries, enabling them to dynamically and appropriately trust the tool responses.

In this paper, we introduce an agentic learning paradigm that enables a multimodal medical agent to empirically learn tools' practical reliability across various types of queries through active interaction with them, thereby learning to trust the correct tool when their outputs conflict. As a concrete instantiation, we study visual question answering (VQA) for chest X-rays (CXRs) and develop a tool-expertise-aware chest X-ray agent (TEA-CXA). Our framework employs reinforcement learning (RL) [15, 23] to allow the MLLM to experimentally accept or reject multimodal tool outputs when they disagree, obtain reward scores, and learn which tool to trust for different types of multimodal queries, as shown in Fig. 1(c). As shown in Fig. 1(d), after training, TEA-CXA acquires awareness of each tool's real-world expertise, and can select the appropriate tool response to trust, surpassing the performance of any single tool[1]. Existing general-purpose tool-using RL frameworks are designed for text-only inputs [3, 21]. We extend and refine these frameworks to support the interaction between MLLM and multimodal tools. Furthermore, our enhanced codebase supports multiple tool calls per turn, parallel tool inference, and image selection for tool invocation when multiple images are in a single user query, which are capabilities well-tailored to medical scenarios. The resulting code framework is flexible and applicable to general scenarios involving RL for multi-turn tool-calling in multimodal contexts.

Our contributions are summarized as follows. (1) Pioneering the consideration of tools' real-world trustworthiness to resolve conflicts among tool responses, moving beyond reliance solely on tools' functional descriptions or pre-specified tool-use traces. (2) Proposing to enable the agent to empirically learn tools' practical reliability across different query types via multimodal agentic learning. (3) Designing a robust code framework for multimodal agentic learning, which is applicable to general multi-turn tool-calling RL in multimodal contexts. (4) Evaluation of TEA-CXA on chest X-ray VQA datasets, demonstrating the superiority of our approach over state-of-the-art methods and a variety of baselines.

2 Method

Fig. 2 illustrates our tool-expertise-aware chest X-ray agent (TEA-CXA) framework via multimodal agentic reinforcement learning. In Sec. 2.1, we introduce tool-expertise-awareness training, which informs the agent about tools' real-world trustworthiness across different types of multimodal queries. In Sec. 2.2, we present our agentic reinforcement learning codebase design that allows an

[1] Note that our method is designed for general medical agents where multiple tools with diverse functionalities are *dynamically* (e.g., varying tool types and arguments) invoked. In our experiments, we use two VQA tools to make the RL training affordable (since training time increases markedly with the number of tools) and to facilitate clear evaluation of tool-response selection performance. Scaling TEA-CXA to an agent with a larger tool suite is left for future work.

MLLM to interact with multimodal tools and incorporates several practically useful enhancements tailored to medical scenarios.

2.1 Tool-expertise-awareness training via multimodal agentic reinforcement learning with multi-turn tool calling

Tools for medical image analysis are inherently error-prone and can produce contradictory outputs. Relying solely on functional descriptions of tools [6] or pre-specified tool-use traces fails to adequately inform an MLLM about the tools' real-world performance and reliability, and thus does not help it decide which tool output to trust when their responses conflict. To bridge this gap, we propose tool-expertise-awareness training, where we leverage reinforcement learning (RL) to enable the agent to actively interact with tools to learn their real-world trustworthiness for various types of multimodal queries.

Reinforcement learning. We adopt Group Relative Policy Optimization (GRPO) [20] as the RL algorithm. Specifically, for each input prompt x, GRPO samples a group of trajectories $\{\tau_1, \tau_2, \ldots, \tau_G\}$ from the reference policy π_{ref}. Then the advantage within the group $\hat{A}_{i,t}$ is computed as:

$$\hat{A}_{i,t} = \frac{R_\phi(\tau) - \text{mean}(\{R_\phi(\tau_1), \ldots, R_\phi(\tau_G)\})}{\text{std}(\{R_\phi(\tau_1), \ldots, R_\phi(\tau_G)\})}, \tag{1}$$

where R_ϕ is the reward function. The policy model is then optimized by maximizing the following objective function:

$$J(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\tau_i|}\sum_{t=1}^{|\tau_i|}\min\left[r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{i,t}\right] - \beta \mathrm{D}_{\text{KL}}\left(\pi_\theta \,\|\, \pi_{\text{ref}}\right), \tag{2}$$

where π_θ is the policy MLLM, $\mathbb{D}_{\text{KL}}$ is the KL divergence, and $r_{i,t}$ is the policy ratio $r_{i,t}(\theta) = \frac{\pi_\theta(\tau_{i,(t)}|\tau_{i,<t})}{\pi_{\text{old}}(\tau_{i,(t)}|\tau_{i,<t})}$.

Tool-expertise-awareness enhancement via multimodal agentic learning with tool calling. We aim to boost the MLLM's awareness of tools' practical reliability across different types of multimodal queries, and enable it to dynamically and correctly select the tool response to trust when tool results contradict. Our framework lets the agent call multiple tools, receive their responses, experimentally trust one of the tool outputs if they disagree, and receive reward scores. In this way, the agent obtains a sense of the tools' realistic accuracies for different types of user queries and thus can select appropriate tools to trust.

As illustrated in Fig. 2, the rollout τ in this paper is the concatenation of interleaved MLLM-generated tool-calling tokens and tool responses. We use `<tool_call>` and `</tool_call>` to enclose a tool-calling action that is written as a JSON object, which will be parsed and executed by the tools, with the tool execution results enclosed in `<tool_response>` and `</tool_response>`. Then, the existing rollout concatenated with the tool responses will be used as the input to generate the MLLM response in the next turn. The rollout process stops when the MLLM outputs the final answer, which is wrapped in `<answer></answer>`.

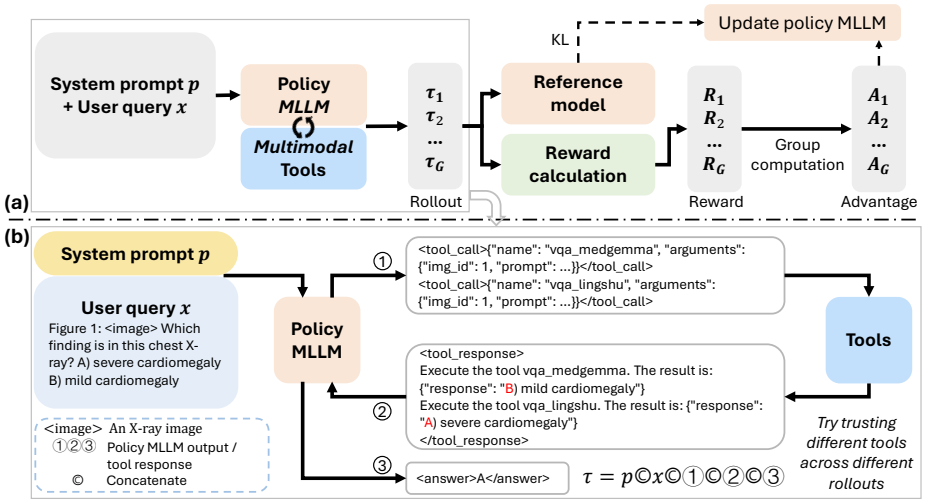

Fig. 2. Overview of the proposed tool-expertise-aware chest X-ray agent (TEA-CXA) framework. (a) The agentic reinforcement learning pipeline with multimodal policy model and multimodal tools. (b) Details of the generation process for a *single* rollout. Tool invocations are dynamically generated by MLLM. Different rollouts try out trusting different tools when tool results contradict.

We design a system prompt that guides the policy MLLM to produce rollouts in the defined format, as demonstrated in Fig. 3(a). In this prompt, the MLLM is instructed to experimentally and randomly trust one tool if tool responses conflict. This prevents the MLLM from biasing toward a specific tool (e.g., one that provides more detailed analysis), given that each tool may have its strengths and weaknesses on different types of queries.

We use a tool-calling format reward $R_t(\tau)$ to verify whether the MLLM adheres to our prescribed tool-calling format, and an answer format reward $R_a(\tau)$ to check the existence of the `<answer></answer>` tags in the rollout. $R_t(\tau)$ and $R_a(\tau)$ are set to 0.1 if their respective conditions are satisfied, and 0 otherwise. The total reward is the sum of the outcome reward and the two format rewards:

$$R(\tau) = R_o(\tau) + R_a(\tau) + R_t(\tau), \tag{3}$$

where the outcome reward $R_o(\tau)$ evaluates the final answer using exact matching. $R_o(\tau)$ is 1 if the answer is correct, and 0 otherwise.

The rollout sequence consists of both the MLLM-generated tokens and tool responses. We apply loss masking for the tool responses [13], ensuring the policy objective gradient is computed only over MLLM-generated tokens.

2.2 Multimodal agentic reinforcement learning framework design

We extend existing agentic training codebases focused on text [3, 21] to robustly support multimodal contexts, with enhancements tailored for medical scenarios.

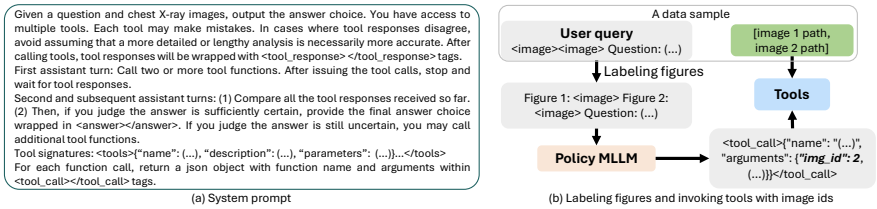

Fig. 3. (a) The system prompt instructs the policy model to follow the specified output format and consider each tool's output as potentially trustworthy when tool results contradict. (b) For queries containing multiple images, our approach directs the policy model to generate image indices instead of file paths for efficient tool invocation. Some details are abbreviated as "(...)" due to space constraints.

The framework is extensible to other domains beyond X-ray analysis, providing a solid base for future research on medical tool-use agents.

Interaction between MLLM and multimodal tools in RL. Existing code frameworks for reinforcement learning with tool calling, e.g., RL-Factory [3], center on LLM policy models and text-only tools, lacking verifiable multimodal support. Building on RL-Factory [3], we expand and refine functionality to enable reliable interaction between MLLM policy model and multimodal tools.

Multiple tool calls per turn and parallel tool inference. Given the imperfect accuracy of medical AI tools, it is beneficial to consult more than one tool in a single turn before proceeding to the next step. We extend [3] to allow for multiple tool calls per turn, each enclosed in `<tool_call>` and `</tool_call>`. Tool results are concatenated and returned to the MLLM. In addition, some medical AI tools have large numbers of parameters and slow inference, becoming the training time bottleneck. To speed up tool responses and accelerate training, our code framework deploys multiple instances of server APIs of the same tool on different ports and performs port round-robin when making calls.

Accommodation for multi-image queries. Although we validate our method on single-image queries, our code framework is readily applicable to multi-image queries. A single query may contain more than one image (e.g., CXRs from different views, including AP, PA, Lateral), while some medical tools accept only one or few images [19]. To allow for image selection when calling tools, we label images in the chat template (e.g., "Figure 1", "Figure 2") and expose the image label as a tool argument. All image paths of a data sample are passed implicitly as a hidden argument. Fig. 3(b) illustrates this workflow. The benefits are two-fold. First, MLLMs can reliably produce short labels, avoiding errors while generating long path strings. Second, this reduces token overhead. The multi-image functionality is verified on ReXVQA [18], though quantitative results are not included due to space constraints.

Table 1. Accuracies (%) on CheXbench, which consists of three subsets: Rad-Restruct, SLAKE, and OpenI. MedRAX denotes its original implementation; "MedRAX*" denotes using the method in MedRAX but applying the same policy MLLM and tools as ours. Note that MedRAX reports mean per-class accuracy, while we use overall accuracy (also referred to as sample-wise accuracy) as metric.

Method	**Overall**	Rad-Restruct	SLAKE	OpenI
Qwen2.5-VL-7B-Instruct [26]	52.4	55.7	70.7	45.5
MedGemma-4B [19]	66.7	60.0	78.9	64.7
Lingshu-7B [28]	70.2	65.2	92.7	64.5
Agent-ensemble	69.3	67.0	87.8	63.9
Reasoning [10]	62.8	65.2	91.9	52.6
Ensemble-A	67.6	60.9	88.6	62.9
Ensemble-B	69.7	61.7	94.3	64.2
CheXagent [4]	62.4	57.1	78.1	59.0
GPT-4o	58.4	53.9	85.4	51.1
MedRAX [6]	61.6	68.7	82.9	52.6
MedRAX* [6]	69.6	61.7	90.2	65.3
TEA-CXA (Ours)	**73.8**	**69.6**	**95.9**	**67.9**

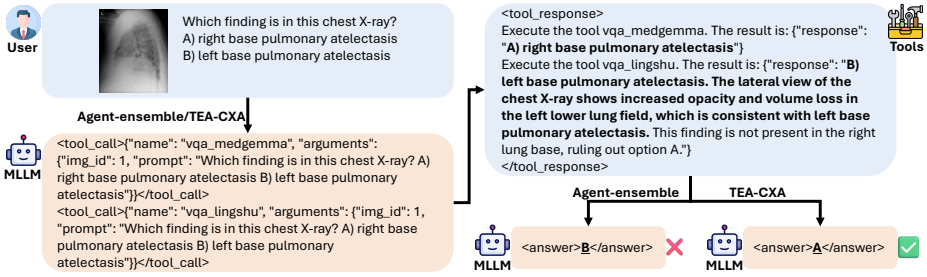

Fig. 4. Qualitative comparison on a sample in CheXbench. Although Lingshu offers more detailed justifications, our method correctly trusts MedGemma's answer, thanks to its awareness of tools' realistic reliability across multimodal query types.

3 Experiments

Datasets. We evaluate TEA-CXA on CheXbench [4] for the visual question answering (VQA) task. We follow MedRAX [6] to use the three subsets of CheXbench: Rad-Restruct, SLAKE, and OpenI, comprising a total of 618 multiple-choice questions. We conduct three-fold cross-validation experiments on CheXbench and ensemble the results across the three folds. Model performance is measured by accuracy (percentage of correct answers).

Implementations. TEA-CXA employs Qwen2.5-VL-7B-Instruct [26] as its policy MLLM. We equip the agent with two high-performance VQA tools: MedGemma-4B [19] and Lingshu-7B [28], to assess the performance gain of our approach with up-to-date tools. Each tool is implemented according to its official HuggingFace

Table 2. Tool response selection accuracies (%) for samples where tool outputs conflict and at least one output is correct.

Method	**Overall**	Rad-Restruct	SLAKE	OpenI
Ensemble-A	46.6	42.3	62.1	43.7
Ensemble-B	54.6	46.2	86.2	48.7
Agent-ensemble	54.0	57.7	58.6	52.1
MedRAX* [6]	57.5	42.3	75.9	55.5
TEA-CXA (Ours)	**63.8**	**69.2**	**89.7**	**56.3**

guidelines. Training is conducted on 8 NVIDIA H800 GPUs. We train TEA-CXA for 100 steps with a batch size of 80. More details are available in our code.

Quantitative performance. We compare our method against the following baselines and models: (1) Direct inference on Qwen2.5-VL-7B-Instruct [26] (our policy MLLM), MedGemma [19], or Lingshu [28] (our tools); (2) Agent-ensemble: invoking both tools and ensembling their outputs with the same policy MLLM, which acts as a naïve agent strategy; (3) Reasoning: training the same policy MLLM to reason and provide the answer with RL [10]. (4) Policy MLLM-tool ensembles: Ensemble-A, which combines the results of Qwen2.5-VL-7B-Instruct and tools, and Ensemble-B, which combines the RL-trained reasoning policy MLLM from (3) with tool outputs. (5) State-of-the-art (SOTA) methods, including biomedical vision-language model CheXagent [4], general-purpose MLLM GPT-4o, and CXR-focused agent MedRAX [6]. The results are shown in Table 1, TEA-CXA outperforms all baselines and prior SOTA.

Qualitative performance. We present one representative case in Fig. 4 comparing TEA-CXA to Agent-ensemble (invoking both tools and ensembling their outputs). The two tools provide inconsistent suggestions. Agent-ensemble incorrectly favors Lingshu, likely due to its more detailed analysis. TEA-CXA correctly trusts MedGemma despite its concise conclusion without analysis. This demonstrates TEA-CXA's ability to resolve conflicting tool outputs based on its awareness of tools' real-world trustworthiness across different multimodal query types instead of surface features of tool outputs.

Tool response selection analysis. Table 2 presents the accuracies of choosing the correct tool response among conflicting outputs when at least one is correct. The compared methods have no prior knowledge about the tools' real-world performance and therefore can only rely on superficial features in the tool outputs to resolve conflicts. In contrast, our approach, trained via multimodal agentic reinforcement learning to internalize each tool's reliability across different query types, selects the correct tool output with substantially higher accuracy.

4 Conclusion

This work presents a novel tool-expertise-aware chest X-ray agent. The agent empirically learns tools' practical trustworthiness via agentic learning. When tool outputs conflict, It experimentally trusts one tool result, receives rewards,

and then learns which tool to trust for each query type. To implement our method, we develop a code framework for multimodal agentic learning with enhancements for medical scenarios, building a foundation for future research. Experiments show that our method outperforms the state-of-the-art methods.

Acknowledgments. This work was supported by a research grant from the Joint Research Scheme (JRS) under the National Natural Science Foundation of China (NSFC) and the Research Grants Council (RGC) of Hong Kong (Project No. N_HKUST654/24) and a research grant from the Innovation and Technology Commission (ITC) (Project No. GHP/124/22).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arora, R.K., Wei, J., Hicks, R.S., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., Sharman, M., Shah, M., Vallone, A., Beutel, A., et al.: Healthbench: Evaluating large language models towards improved human health. arXiv preprint arXiv:2505.08775 (2025)
2. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., Meissen, F., Ranjit, M.P., Srivastav, S., Gong, J., Codella, N.C.F., Falck, F., Oktay, O., Lungren, M.P., Wetscherek, M.T.A., Alvarez-Valle, J., Hyland, S.L.: Maira-2: Grounded radiology report generation. arXiv **abs/2406.04449** (2024), https://arxiv.org/abs/2406.04449
3. Chai, J., Yin, G., Xu, Z., Yue, C., Jia, Y., Xia, S., Wang, X., Jiang, J., Li, X., Dong, C., et al.: Rlfactory: A plug-and-play reinforcement learning post-training framework for llm multi-turn tool-use. arXiv preprint arXiv:2509.06980 (2025)
4. Chen, Z., Varma, M., Delbrouck, J.B., Paschali, M., Blankemeier, L., Van Veen, D., Valanarasu, J.M.J., Youssef, A., Cohen, J.P., Reis, E.P., et al.: Chexagent: Towards a foundation model for chest x-ray interpretation. In: AAAI 2024 Spring Symposium on Clinical Foundation Models (2024)
5. Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M., et al.: Torchxrayvision: A library of chest x-ray datasets and models. In: International Conference on Medical Imaging with Deep Learning. pp. 231–249. PMLR (2022)
6. Fallahpour, A., Ma, J., Munim, A., Lyu, H., WANG, B.: MedRAX: Medical reasoning agent for chest x-ray. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=JiFfij5iv0
7. Ferber, D., El Nahhas, O.S., Wölflein, G., Wiest, I.C., Clusmann, J., Leßmann, M.E., Foersch, S., Lammert, J., Tschochohei, M., Jäger, D., et al.: Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. Nature cancer pp. 1–13 (2025)
8. Gao, S., Zhu, R., Kong, Z., Noori, A., Su, X., Ginder, C., Tsiligkaridis, T., Zitnik, M.: Txagent: An ai agent for therapeutic reasoning across a universe of tools. arXiv preprint arXiv:2503.10970 (2025)

9. Ghezloo, F., Seyfioglu, M.S., Soraki, R., Ikezogwo, W.O., Li, B., Vivekanandan, T., Elmore, J.G., Krishna, R., Shapiro, L.: Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. arXiv preprint arXiv:2502.08916 (2025)
10. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
11. He, Y., Li, A., Liu, B., Yao, Z., He, Y.: Medorch: Medical diagnosis with tool-augmented reasoning agents for flexible extensibility. arXiv preprint arXiv:2506.00235 (2025)
12. Hoopes, A., Butoi, V., Guttag, J., Dalca, A.: Voxelprompt: A vision-language agent for grounded medical image analysis. oct. 10, 2024, arxiv. arXiv preprint arXiv:2410.08397
13. Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H., Han, J.: Search-r1: Training llms to reason and leverage search engines with reinforcement learning. arXiv preprint arXiv:2503.09516 (2025)
14. Jin, Q., Wang, Z., Yang, Y., Zhu, Q., Wright, D., Huang, T., Khandekar, N., Wan, N., Ai, X., Wilbur, W.J., et al.: Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. Nature Communications **16**(1), 9377 (2025)
15. Kaelbling, L.P., Littman, M.L., Moore, A.W.: Reinforcement learning: A survey. Journal of artificial intelligence research **4**, 237–285 (1996)
16. Li, B., Yan, T., Pan, Y., Luo, J., Ji, R., Ding, J., Xu, Z., Liu, S., Dong, H., Lin, Z., et al.: Mmedagent: Learning to use medical tools with multi-modal agent. arXiv preprint arXiv:2407.02483 (2024)
17. Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., et al.: Agentbench: Evaluating llms as agents. arXiv preprint arXiv:2308.03688 (2023)
18. Pal, A., Lee, J.O., Zhang, X., Sankarasubbu, M., Roh, S., Kim, W.J., Lee, M., Rajpurkar, P.: Rexvqa: A large-scale visual question answering benchmark for generalist chest x-ray understanding. arxiv (2025)
19. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
20. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
21. Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., Wu, C.: Hybridflow: A flexible and efficient rlhf framework. arXiv preprint arXiv: 2409.19256 (2024)
22. Sun, Y., Si, Y., Zhu, C., Zhang, K., Shui, Z., Ding, B., Lin, T., Yang, L.: Cpathagent: An agent-based foundation model for interpretable high-resolution pathology image analysis mimicking pathologists' diagnostic logic. arXiv preprint arXiv:2505.20510 (2025)
23. Sutton, R.S., Barto, A.G., et al.: Reinforcement learning. Journal of Cognitive Neuroscience **11**(1), 126–134 (1999)
24. Tang, X., Shao, D., Sohn, J., Chen, J., Zhang, J., Xiang, J., Wu, F., Zhao, Y., Wu, C., Shi, W., et al.: Medagentsbench: Benchmarking thinking models and agent frameworks for complex medical reasoning. arXiv preprint arXiv:2503.07459 (2025)

25. Tanno, R., Barrett, D.G., Sellergren, A., Ghaisas, S., Dathathri, S., See, A., Welbl, J., Lau, C., Tu, T., Azizi, S., et al.: Collaboration between clinicians and vision–language models in radiology report generation. Nature Medicine **31**(2), 599–608 (2025)
26. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
27. Xie, J., Chen, Z., Zhang, R., Wan, X., Li, G.: Large multimodal agents: A survey. arXiv preprint arXiv:2402.15116 (2024)
28. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
29. Zhao, W., Wu, C., Fan, Y., Zhang, X., Qiu, P., Sun, Y., Zhou, X., Wang, Y., Sun, X., Zhang, Y., et al.: An agentic system for rare disease diagnosis with traceable reasoning. arXiv preprint arXiv:2506.20430 (2025)