# MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision-Language Models (VLMs) via Reinforcement Learning

Jiazhen Pan[1,2*], Che Liu[3*], Junde Wu[2], Fenglin Liu[2], Jiayuan Zhu[2], Hongwei Bran Li[4], Chen Chen[5,6], Cheng Ouyang[2,6†], Daniel Rueckert[1,6†]

[1] Chair for AI in Healthcare and Medicine, Technical University of Munich (TUM) and TUM University Hospital, Germany
[2] Department of Engineering Science, University of Oxford, UK
[3] Data Science Institute, Imperial College London, UK
[4] Massachusetts General Hospital, Harvard Medical School, USA
[5] School of Computer Science, University of Sheffield, UK
[6] Department of Computing, Imperial College London, UK
jiazhen.pan@tum.de, che.liu21@imperial.ac.uk

**Abstract.** Reasoning is a critical frontier for advancing medical image analysis, where transparency and trustworthiness play a central role in both clinician trust and regulatory approval. Although Medical Visual Language Models (VLMs) show promise for radiological tasks, most existing VLMs merely produce final answers without revealing the underlying reasoning. To address this gap, we introduce MedVLM-R1, a medical VLM that explicitly generates natural language reasoning to enhance transparency and trustworthiness. Instead of relying on supervised fine-tuning (SFT), which often suffers from overfitting to training distributions and fails to foster genuine reasoning, MedVLM-R1 employs a reinforcement learning framework that incentivizes the model to discover human-interpretable reasoning paths without using any reasoning references. Despite limited training data (600 visual question answering samples) and model parameters (2B), MedVLM-R1 boosts accuracy from 55.11% to 78.22% across MRI, CT, and X-ray benchmarks, outperforming larger models trained on over a million samples. It also demonstrates robust domain generalization under out-of-distribution tasks. By unifying medical image analysis with explicit reasoning, MedVLM-R1 marks a pivotal step toward trustworthy and interpretable AI in clinical practice.

**Keywords:** Medical reasoning · Reinforcement learning · VLMs

## 1 Introduction

Radiological images are fundamental to modern healthcare, with over 8 billion scans performed annually [2]. As diagnostic demand grows, the demand for

* Equal contribution † Equal advice
Code is available at https://github.com/JZPeterPan/MedVLM-R1.
Inference model is available at https://huggingface.co/JZPeterPan/MedVLM-R1.

efficient AI-driven interpretation becomes increasingly acute. Medical Vision-Language Models (VLMs), developed for radiological visual question answering (VQA) in MRI, CT and X-ray images, offer substantial promise in assisting clinicians/patients. Recent advances in general-purpose LLMs/VLMs (e.g., GPT-4o [16], Claude-3.7 Sonnet [3]) highlight sophisticated reasoning capabilities. However, the medical domain places an especially high premium on explainable decision-making: both clinicians/patients need to understand not just *what* conclusion was reached, but also *why*. Existing medical VLMs often provide only final answers or "quasi-explanations" derived from pre-training pattern matching, which do not necessarily reflect genuine, step-by-step reasoning. Consequently, ensuring interpretability and trustworthiness remains an urgent challenge in real-world clinical settings.

We argue that the limited reasoning capability for existing medical VLM is primarily due to the inherent drawbacks of Supervised Fine-Tuning (SFT) [1, 25, 9] which is the most common strategy for adapting large foundation models for specialized medical tasks [13, 21, 6]. Despite its simplicity, SFT faces two critical challenges: 1) An over-reliance on final-answer supervision often leads to overfitting, shortcut learning, and weaker performance on out-of-distribution (OOD) data – an issue particularly consequential in high-stake medical scenarios [11]. 2) Direct supervision with only final answers provides minimal incentive for cultivating reasoning abilities within VLMs. A possible mitigation is distilling a more capable teacher model's chain-of-thought (CoT) reasoning for SFT [31, 20]. However, constructing high-quality CoT data is prohibitively expensive to scale in specialized domains like healthcare. As a result, current medical VLMs that rely on SFT often fall short of delivering transparent explanations and robust generalizations when confronted with unfamiliar data.

In contrast, Reinforcement Learning (RL) [26] offers a compelling alternative for cultivating emergent reasoning by rewarding models for discovering their own logical steps rather than memorizing final answers or copying teacher CoT rationales. Indeed, a recent work ***SFT Memorizes, RL Generalizes*** [11] confirms that RL-trained models often display superior generalization compared to their SFT counterparts. However, conventional RL pipelines typically depend on auxiliary neural reward models, requiring substantial resources to continuously update both policy and reward models [35, 24]. A promising alternative, group relative policy optimization (GRPO) [27], eliminates the need for neural reward models by employing a rule-based group-relative advantage strategy (see sec. 3 for more details). This approach has demonstrated advanced reasoning, fostering capabilities while reducing computational demands in DeepSeek-R1 [12]. Despite its potential benefits for resource- and data-constrained domains like healthcare, GRPO remains largely unexplored in medical contexts.

In this work, we introduce **MedVLM-R1**, the first medical VLM capable of generating answers with explicit reasoning by training with GRPO for radiology VQA tasks. Our current scope specifically focuses on VQA tasks. Our contributions are as follows:

1. **Medical VLM with Explicit Reasoning**: We introduce MedVLM-R1, the first lightweight medical VLM capable of generating explicit reasoning alongside the final answer, rather than providing only the final answer.
2. **Emerging Reasoning Without Explicit Supervision**: Unlike traditional SFT methods that require data with complex reasoning steps, MedVLM-R1 is trained using GRPO with datasets containing only final answers, demonstrating emergent reasoning capabilities without explicit supervision.
3. **Superior Generalization and Efficiency**: MedVLM-R1 achieves robust generalization to out-of-distribution data (e.g. MRI $\rightarrow$ CT/X-ray) and outperforms larger models like Qwen2VL-72B and Huatuo-GPT-Vision-7B, despite being a compact 2B-parameter model trained on just 600 samples.

## 2 Related Work

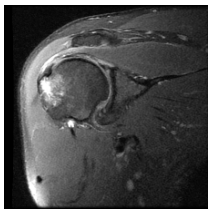
***Medical VLMs and Their Limitations.*** The rise of large-scale VLMs has spurred numerous domain-specific adaptations for healthcare, with systems such as LLaVA-Med [19] and HuatuoGPT-Vision [7] achieving impressive results in radiology VQA and related diagnostic tasks. Despite these advancements, using SFT on final-answer labels remains the dominant strategy for tailoring large models to medical domains [34, 6, 33, 5]. This approach generally requires substantial amounts of high-quality image-text data (ranging from 660k [19] to 32M samples [32]) which is costly to curate and often hampered by noise/privacy concerns. Moreover, the reliance on final-answer supervision provides limited scope for exposing a model's intermediate reasoning—an important factor in building clinicians' trust. In addition, SFT-based models often overfit to narrow training distributions, leading to weaker generalization on OOD clinical scenarios.

***Reinforcement Learning for Enhanced Reasoning.*** To mitigate SFT's limitations, RL [29, 10, 35, 24, 17] has emerged as a compelling alternative for improving model interpretability and robustness. Classic RL methods, such as proximal policy optimization (PPO) [26], have been widely adopted in text-based learning (*e.g.,* policy shaping for LLMs) and can reward not only correctness but also the quality of intermediate reasoning steps. Recent studies suggest that while SFT "memorizes," RL can help models "generalize" [11], offering a more stable trajectory toward domain-transferable representations. Notably, GRPO [27] extends PPO by eliminating its (neural) value function estimator and focusing on a rule-based group-relative advantage for selecting actions, showing promise in resource-constrained settings like DeepSeek-R1 [12]. Such RL-driven frameworks could be particularly beneficial for medical tasks, where limited data availability, high-stakes decision-making, and the need for explicit reasoning converge. In the following section, we will detail how these insights motivate our approach.

## 3 Methods

***Overview.*** We leverage RL to incentivize explicit reasoning capabilities in medical VLMs, specifically employing GRPO due to its efficiency and effectiveness.

**Question Template**



What can be observed in this image?
A) Cartilage degeneration B) Labral pathology
C) Bone fracture D) Tendonitis

**Model Input Prompt**

{Question} Your task:
1. Think through the question step by step, enclose your reasoning process in <think>...</think> tags.
2. Then provide the correct single-letter choice (A, B, C, D,...) inside <answer>...</answer> tags.
3. No extra information or text outside of these tags.

**Model Output Examples**

<think> The image is a grayscale MRI image of an upper arm joint. The bicondylar humeral head of the humerus is visible. There is a well-defined ... </think>
<answer>B, there is no clear indication of ... </answer>

**Format Reward** = 1 due to the present of all tags and no content outside the tags
**Accuracy Reward** = 0.5 due to extra explanation appended after the answer

Fig. 1: The template of our employed prompt, an example of model's response and reward criterion.

While the seminal work [27] applies GRPO to reasoning in coding and mathematics, we adapt these principles to the medical domain, specifically radiology data (MRI, CT, X-ray). Our approach incorporates medical imaging prompts and custom reward functions designed to encourage explicit reasoning and domain-specific answer formats. This work serves as an initial exploration of using pure reinforcement learning to build a multi-modal medical reasoning model.

***Base Model and Prompt Template.*** We adopt a state-of-the-art VLM – Qwen2-VL-2B [30] as our base model, denoted by $\pi_\theta$, where $\theta$ are the trainable parameters. Given a training dataset $\boldsymbol{V}$, each sample $\boldsymbol{v}$ consists of: 1) An image $f$, which is a radiology image and 2) a text prompt $q$, composed of the user's question alongside a fixed system message, as illustrated in Figure 1. The VLM then produces an output $\{o\}$, which includes both a reasoning trace and a final answer in designated XML-like tags (`<think>...</think>` and `<answer>...</answer>`). Our RL objective is to optimize $\pi_\theta$ so that answers are accurate, well-formatted, and provide transparent reasoning.

***Group Relative Policy Optimization (GRPO).*** To encourage robust, interpretable responses, we employ GRPO [27], an RL algorithm that extends PPO by focusing on a group-relative advantage instead of a learned value function. Concretely, at each training step:

1. We sample $G$ candidate outputs $\{o_i\}_{i=1}^{G}$ from $\pi_{\theta_{old}}$, the model parameters before the current update.

2. We compute a reward $r_i$ for each output using a reward function (see next paragraph). Based on $r_i$ we calculate a group relative advantage $A_i$ which is normalized by the group statistics: $A_i = \frac{r_i - mean(\{r_1, r_2, \cdots, r_G\})}{std(\{r_1, r_2, \cdots, r_G\})}$. A reward above the group average is advantaged and can further incentivize the model.
3. Our VLM model $\pi_\theta$ is then updated by maximizing $\mathcal{J}_{GRPO}$ which incorporates a clipped regularization on the relative advantage estimation for model preference alignment and training stability:

$$\begin{aligned} &\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{\boldsymbol{v}\sim P(\boldsymbol{V})}\mathbb{E}_{\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(\cdot|\boldsymbol{v})} \\ &\frac{1}{G}\sum_{i=1}^{G}\Big[\min\Big(r_i^{\text{ratio}}A_i, \text{clip}\left(r_i^{\text{ratio}}, 1\pm\epsilon\right)A_i\Big) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\Big] \end{aligned} \tag{3.1}$$

with $r_i^{\text{ratio}} = \frac{\pi_\theta(o_i|\boldsymbol{v})}{\pi_{\theta_{\text{old}}}(o_i|\boldsymbol{v})}$. An additional *Kullback–Leibler* term $\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{ref})$ is applied to penalize divergence from a reference model $\pi_{\text{ref}}$ (the initial checkpoint), helping prevent catastrophic forgetting. $\epsilon, \beta \in \mathbb{R} \geq 0$ control the regularization strengths.

***Reward function.*** Specifically, for multiple-choice medical VQA tasks, we propose a two-part rule-based reward function, inspired by [12]:
1) *Format Reward.* We incentivize outputs that provide a reasoning trace within the tags `<think>` ... `</think>` and a succinct final answer within the tag `<answer>` ... `</answer>`. If all four tags are present exactly once and no content is present outside these tags, we assign a format reward of 1. Any missing/duplicated tags or content outside yield 0.
2) *Accuracy Reward.* After verifying the correct format, we evaluate the correctness of the final answer. Specifically, If the letter choice `A, B, C, D,...` inside the `<answer>` ... `</answer>` tag and it responds with the ground-truth choice exactly, that is an exact match with a reward of 1 point. Further, if the letter is correct but contains additional explanations or uses the choice content instead of the corresponding letter (e.g., "A: Pulmonary nodule" or "Pulmonary nodule"), that is a partial match with a 0.5 points reward. However, if the letter does not match, is missing, or the answer is not enclosed in the answer tag, that is an incorrect or missing answer and no reward would be granted (0 points).

The total reward $r_i \in [0, 2]$ is the sum of both format and accuracy reward. By structuring the reward function in this hierarchy (format before correctness), we guide the model to first adopt the desired response structure, then refine its answer selection for accurate, interpretable medical reasoning. It is worth noting that both terms are necessary since without *Format Reward*, the final answer cannot be extracted while without *Accuracy Reward*, the model cannot converge.

## 4 Experiments

***Dataset.*** We conduct our experiments using the HuatuoGPT-Vision [7] evaluation dataset, which is a processed and combined dataset from several publicly

available medical VQA benchmarks, including VQA-RAD [18], SLAKE [22], PathVQA [14], OmniMedVQA [15], and PMC-VQA [34]. In total, the dataset comprises 17,300 multiple-choice questions linked to images covering various medical imaging modalities, with 2–6 possible choices per question. For this study, we focus on radiology modalities: CT, MRI, and X-ray. Specifically, we use 600 MRI image-question pairs for training and set aside 300 MRI, 300 CT, and 300 X-ray pairs for testing. The MRI test set is used as in-domain test, whereas the CT and X-ray test set serve as OOD test.

***Implementation details.*** We adopt Qwen2-VL-2B as our base VLM. This model is originally trained on data from curated web pages, open-source datasets, and synthetic sources. To adapt it to the medical domain, we employ the GRPO reinforcement learning framework outlined in Section 3. Our implementation builds on the public VLM reasoning repositories [23, 8, 28]. We perform fine-tuning on two NVIDIA A100 SXM4 80GB for 300 steps, using a batch size of 2, which takes approximately 4 hours. Generation candidate number $G$ is set to 6. The other training optimization hyper-parameters are set as suggested by [8].

***Baseline methods and evaluation metric.*** We compare MedVLM-R1 with the following baselines: 1. Qwen2-VL family [4] including Qwen2-VL-2B (the unmodified base model), Qwen2-VL-7B and -72B which are the large/huge model variants. 2. HuatuoGPT-vision [7]: A medical VLM built upon Qwen2-VL-7B. 3. SFT: The same Qwen2-VL-2B base model fine-tuned with standard SFT, using the same training setting with 600 MRI question-answer pairs. We apply negative log-likelihood as the loss function to carry out the SFT training. All baselines use a simple prompting format, *e.g.,* `{Question} Your task: provide the correct single-letter choice (A, B, C, D, ...).` In contrast, MedVLM-R1 uses the RL-based prompt as described in Section 3, designed to elicit explicit reasoning. For evaluation, each model receives one point for the correct single-letter answer and zero otherwise. In the test of MedVLM-R1, only the correct choice enclosed in the `<answer>...</answer>` tag is scored as correct; any deviation from this format, even if semantically correct, results in a zero score.

## 5 Results and Discussion

***Overall Performance.*** Table 1 summarizes both in-domain (ID) and out-of-domain (OOD) performance for various VLMs. Note that ID/OOD comparisons specifically refer to models fine-tuned on MRI data. Unsurprisingly, VLMs fine-tuned with both GRPO and SFT significantly outperform zero-shot general-purpose VLMs on in-domain tasks. Our GRPO-trained model shows very strong OOD performance, achieving a **16%** improvement on CT and a **35%** improvement on X-ray compared to SFT counterparts, underscoring GRPO's superior generalizability. Furthermore, despite being a compact 2B-parameter model trained on just 600 samples, MedVLM-R1 outperforms larger models like Qwen2-VL-72B and HuatuoGPT-Vision-7B, with the latter being specifically

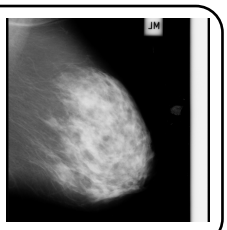
Problem 1: What can be observed in this image?
A) Lungs B) Bladder C) Brain D) Heart

**MedVLM-R1 Output:**

<think>
The image is a chest X-ray, which is a type of radiographic image used to visualize the internal structures of the body, particularly the lungs and bones. The presence of lung markings and the ribcage are characteristic features of a chest X-ray.
</think>

<answer>A</answer> **Groundtruth Answer: A**

Problem 2: What content appears in this image?
A) Cardiac tissue, B) Breast tissue, C) Liver tissue, D) Skin tissue



**MedVLM-R1 Output:**

<think> The image appears to be a mammogram, which is a type of X-ray used to detect abnormalities in the breast tissue. The image shows the breast tissue with various densities and patterns, which are typical of mammograms. </think>

<answer>B</answer> **Groundtruth Answer: B**

Problem 3: What is the finding in this image?
A) Stomach ulcer, B) Nerve compression,
C) Cord pathology, D) Liver cirrhosis

**MedVLM-R1 Output:**

<think> To determine the correct answer, we need to identify the finding in the image provided. The MRI scan in question is likely a view of a spine, not the stomach, or the liver. Stomach ulcers, liver cirrhosis, and spinal cord pathology can be diagnosed with MRI, but the image in question shows a spinal view, not a gastrointestinal or abdominal view. Therefore, the only reasonable options left are nerve compression and cord pathology.
</think>

<answer>C</answer> **Groundtruth Answer: C**

Problem 4: What is the specific abnormality present in this image?
A) Gastritis, B) Arthritis, C) Appendicitis, D) Urolithiasis

**MedVLM-R1 Output:**

<think> The image presented is a transverse CT scan of the abdomen and pelvis. The presence of calculi (urines filled with stones or grit) in the pelvic organs is a consistent finding in urolithiasis. </think>

<answer>D</answer> **Groundtruth Answer: D**

Fig. 2: Medical VQA examples of MedVLM-R1 on X-ray (1, 2), MRI (3) and CT (4).

trained on large-scale medical data. This highlights the immense potential of RL-based training methods for efficient and scalable medical VLM development.

***Reasoning Competence and Interpretability.*** Beyond strong generalization, a central strength of MedVLM-R1 is its ability to produce explicit reasoning—a capability absent in all baselines. As illustrated in Figure 2, MedVLM-R1 presents a logical thought process within the `<think>` tag, with the final decision enclosed in the `<answer>` tag. Notably, for relatively simpler questions (prob-

Table 1: Results of VQA-VLMs on MRI (in-domain), and CT and X-Ray (out-of-domain) modalities. "−2B" indicates the model has 2 billion parameters, etc.

| Method | Num. of Seen Medical Sample | In-Domain / Out-of-Domain | | | Average |
|---|---|---|---|---|---|
| | | (MRI→MRI) | (MRI→CT) | (MRI→X-ray) | |
| Random Guess | / | 25.00 | 30.25 | 26.00 | 27.08 |
| *Zero-shot VLM* | | | | | |
| Qwen2-VL-2B | / | 61.67 | 50.67 | 53.00 | 55.11 |
| Qwen2-VL-7B | / | 72.33 | 68.67 | 66.63 | 69.21 |
| Qwen2-VL-72B | / | 68.67 | 60.67 | 72.33 | 67.22 |
| *Zero-shot Medical VLM* | | | | | |
| Huatuo-GPT-vision-7B | 1,294,062 | 71.00 | 63.00 | **73.66** | 69.22 |
| *MRI fine-tuned VLM* | | | | | |
| Qwen2-VL-2B (SFT) | 600 | 94.00 | 54.33 | 34.00 | 59.44 |
| **Ours-2B (GRPO)** | 600 | **95.33** | **70.33** | 69.00 | **78.22** |

lem 1 and 2), the reasoning appears cogent and aligned with medical knowledge. However, more complex queries sometimes reveal heuristic or just partial reasoning. For example, in the third sample, the model arrives at the correct answer via the **process of elimination** rather than detailed medical analysis, suggesting it leverages cue-based reasoning instead of domain expertise. Likewise, in some instances (e.g., question 4), the causal chain between reasoning and conclusion remains unclear, raising the question of whether the model merely **retrofits an explanation after predicting the correct answer**. Despite these imperfections, MedVLM-R1 represents a notable step toward interpretability in radiological decision-making.

***Limitations.*** Although MedVLM-R1 demonstrates promising results in MRI, CT, and X-ray datasets, several limitations remain: 1. Modality Gaps: When tested on other medical modalities (e.g., pathology or OCT images), the model fails to converge. We hypothesize this arises from the base model's insufficient exposure to such modalities during pre-training. 2. Closed-Set Dependence: The current approach is tailored to multiple-choice (closed-set) VQA. In open-ended question settings where no predefined options are provided, the model's performance degrades substantially. This is also a common challenge for many VLMs. 3. Superficial/hallucinated Reasoning: In some reasoning cases, MedVLM-R1 provides a correct answer without offering a meaningful reasoning process (e.g., `<think>To determine the correct observation from this spine MRI, let's analyze the image.</think><answer>B</answer>`). Additionally, we have not yet analyzed the variability of the model's answers or visualized its attention maps of the vision encoder—both of which are left for future work. Moreover, sometimes the model concludes a correct choice while providing an inference that can lead to another answer. This phenomenon underscores that even models designed for explainability can occasionally revert to superficial/hallucinated justifications, highlighting an ongoing challenge in generating consistently transparent and logically sound rationales. Regarding all these issues, we believe the

current 2B-parameter scale of our base model constitutes a potential bottleneck, and we plan to evaluate MedVLM-R1 on larger VLM backbones to address these concerns. We will also evaluate our model's generalizability on broader medical VLM applications such as grounding and report generation.

## 6 Conclusion

We present MedVLM-R1, a medical VLM that integrates GRPO-based reinforcement learning to bridge the gap between accuracy, interpretability, and robust performance in radiology VQA. By focusing on explicit reasoning, the model fosters transparency and trustworthiness—qualities essential in high-stakes clinical environments. Our results demonstrate that RL-based approaches generalize better than purely SFT methods, particularly under OOD settings. Although VLM-based medical reasoning is still at a nascent stage and faces considerable challenges, we believe that its potential for delivering safer, more transparent AI-driven healthcare solutions will be appreciated and should be encouraged.

***Acknowledgements.*** This work is partially funded by the European Research Council (ERC) project Deep4MI (884622). JW is supported by the Engineering and Physical Sciences Research Council (EPSRC) under grant EP/S024093/1 and GE HealthCare. FL is supported by the Clarendon Fund. JZ is supported by the EPSRC under grant EP/S024093/1 and Global Health R&D of the healthcare business of Merck KGaA, Darmstadt, Germany. HL is supported by a Postdoc Mobility Grant from SNSF. CC is funded by the Royal Society (RGS/R2/242355). CO is supported by UKRI grant EP/X040186/1.

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

## References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Akhter, Y., Singh, R., Vatsa, M.: Ai-based radiodiagnosis using chest x-rays: A review. Frontiers in big data **6**, 1120989 (2023)
3. Anthropic: Claude 3.7 sonnet system card (2025)
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
5. Chaves, J.M.Z., Huang, S.C., Xu, Y., Xu, H., Usuyama, N., Zhang, S., Wang, F., Xie, Y., Khademi, M., Yang, Z., et al.: Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. arXiv preprint arXiv:2403.08002 (2024)
6. Chen, J., Yang, D., Jiang, Y., Li, M., Wei, J., Hou, X., Zhang, L.: Efficiency in focus: Layernorm as a catalyst for fine-tuning medical visual language pre-trained models. arXiv preprint arXiv:2404.16385 (2024)

7. Chen, J., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Zhang, R., Cai, Z., Ji, K., Yu, G., Wan, X., Wang, B.: Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale (2024)
8. Chen, L., Li, L., Zhao, H., Song, Y., Vinci: R1-v: Reinforcing super generalization ability in vision-language models with less than $3. `https://github.com/Deep-Agent/R1-V` (2025), accessed: 2025-02-02
9. Chen, W., Ma, X., Wang, X., Cohen, W.W.: Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. arXiv preprint arXiv:2211.12588 (2022)
10. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. Advances in neural information processing systems **30** (2017)
11. Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q.V., Levine, S., Ma, Y.: Sft memorizes, rl generalizes: A comparative study of foundation model post-training. arXiv preprint arXiv:2501.17161 (2025)
12. 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)
13. Hartsock, I., Rasool, G.: Vision-language models for medical report generation and visual question answering: A review. Frontiers in Artificial Intelligence **7** (2024)
14. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
15. Hu, Y., Li, T., Lu, Q., Shao, W., He, J., Qiao, Y., Luo, P.: Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In: Conference on Computer Vision and Pattern Recognition. pp. 22170–22183 (2024)
16. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
17. Kumar, A., Zhuang, V., Agarwal, R., Su, Y., Co-Reyes, J.D., Singh, A., Baumli, K., Iqbal, S., Bishop, C., Roelofs, R., et al.: Training language models to self-correct via reinforcement learning. arXiv preprint arXiv:2409.12917 (2024)
18. Lau, J.J., Gayen, S., Ben Abacha, A., et al.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)
19. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
20. Li, L.H., Hessel, J., Yu, Y., Ren, X., Chang, K.W., Choi, Y.: Symbolic chain-of-thought distillation: Small models can also" think" step-by-step. arXiv preprint arXiv:2306.14050 (2023)
21. Lian, C., Zhou, H.Y., Yu, Y., Wang, L.: Less could be better: Parameter-efficient fine-tuning advances medical vision foundation models. arXiv preprint arXiv:2401.12215 (2024)
22. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: international symposium on biomedical imaging (ISBI). pp. 1650–1654 (2021)
23. LMMs-Lab: open-r1-multimodal. `https://github.com/EvolvingLMMs-Lab/open-r1-multimodal` (2025), accessed: 2025-01-27
24. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow

instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
25. Qwen-Team: Qwq: Reflect deeply on the boundaries of the unknown (2024)
26. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
27. 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)
28. Shen, H., Zhang, Z., Zhang, Q., Xu, R., Zhao, T.: Vlm-r1: A stable and generalizable r1-style large vision-language model. `https://github.com/om-ai-lab/VLM-R1` (2025), accessed: 2025-02-15
29. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al.: Mastering the game of go without human knowledge. nature **550**(7676), 354–359 (2017)
30. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
31. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
32. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. arXiv preprint arXiv:2308.02463 (2023)
33. Zhang, K., Zhou, R., Adhikarla, E., Yan, Z., Liu, Y., Yu, J., Liu, Z., Chen, X., Davison, B.D., Ren, H., et al.: A generalist vision–language foundation model for diverse biomedical tasks. Nature Medicine pp. 1–13 (2024)
34. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)
35. Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019)