CARE: Confidence-Aware Reasoning for Reliable Medical VQA

Yuetian Du[1†], Yucheng Wang[1†], Zhenyuan Chen[1], Luyuan Chen[1], Rongyu Zhang[1], Jinjian Zhang[2], Wei Zhou[2], Zhijie Xu[3], Ming Kong[1], Zhan Zhou[1], Jie Liu[4✉], and Qiang Zhu[1✉]

[1] Zhejiang University [2] Ant Group
[3] University of Michigan [4] City University of Hong Kong
{22421227,zhuq}@zju.edu.cn

Abstract. Reinforcement Fine-Tuning (RFT) has enabled medical Multimodal Large Language Models (MLLMs) to produce Chain-of-Thought (CoT) reasoning for visual question answering, yet these models suffer from *confidence miscalibration*—a systematic gap between expressed certainty and actual diagnostic accuracy that undermines clinical trust. We propose **CARE**, a **C**onfidence-**A**ware medical **RE**asoning framework that jointly optimizes accuracy and calibration through a dual-stage pipeline. First, a scalable Medical-CoT synthesis provides structured cold-start data for Supervised Fine-Tuning. Second, Group Relative Policy Optimization (GRPO) with a novel **Confidence-Aware Reward (CAR)** mechanism ties the model's confidence to diagnostic correctness within the reward signal. Across three Medical VQA benchmarks, **CARE** achieves the highest diagnostic accuracy while obtaining the lowest Expected Calibration Error and Hallucination Rate, establishing a foundation for trustworthy clinical decision support. Our code is available at https://github.com/anotherbricki/CARE.

Keywords: Reinforcement Fine-Tuning · Medical VQA · Confidence Calibration.

1 Introduction

Medical Multimodal Large Language Models (MLLMs) are increasingly applied to clinical decision support tasks such as medical visual question answering (VQA) [12]. Typically built through Supervised Fine-Tuning (SFT) on curated clinical instruction datasets, these models learn a direct input-output mapping that produces predictions without transparent reasoning processes. In clinical practice, where diagnostic errors carry severe consequences, physicians are unlikely to trust model predictions that lack interpretable reasoning chains. This highlights an urgent need for medical MLLMs that can perform transparent, step-by-step clinical reasoning.

† Equal contribution. ✉ Corresponding author.

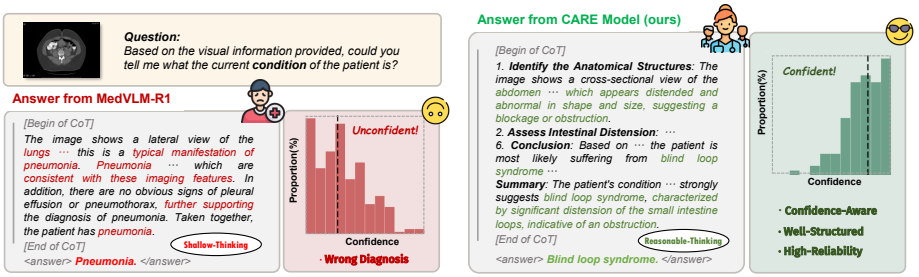

Fig. 1. *Comparative Analysis of Diagnostic Reasoning Paths.* Existing medical reasoning MLLMs such as MedVLM-R1 exhibit confidence miscalibration, where expressed certainty fails to reflect actual diagnostic accuracy. **CARE** explicitly aligns subjective confidence with diagnostic correctness, producing verifiable and interpretable reasoning trajectories for clinical decision-making.

Reinforcement Fine-Tuning (RFT), where RL algorithms such as Group Relative Policy Optimization (GRPO) [16] optimize models against verifiable reward signals, has recently enabled reasoning-focused models that generate extended Chain-of-Thought (CoT) outputs. This paradigm has been adopted in the medical domain, producing early medical reasoning models [9,15] for Medical VQA. While these models improve reasoning transparency, a critical issue persists: *confidence miscalibration* [21,20,5]. As illustrated in Figure 1, existing models exhibit a systematic misalignment between expressed confidence and actual diagnostic accuracy: they may appear underconfident on correct diagnoses or assign unwarranted certainty to incorrect ones. In clinical deployment, such miscalibration undermines trust, as a model that cannot reliably reflect its own uncertainty provides no safe basis for decision-making.

Related Work. Reinforcement Fine-Tuning (RFT) has emerged as a new paradigm for improving reasoning in large models [13,22], surpassing traditional SFT-based CoT approaches. Unlike SFT, which passively fits existing data distributions, RFT enables models to receive direct feedback from verifiable outcomes and dynamically refine their reasoning paths [2]. Pioneered in mathematical reasoning [16] and popularized by DeepSeek-R1 [3], RFT has recently been extended to multimodal medical applications, producing models such as Med-R1 [9], MedVLM-R1 [15], and others [18,4,19,8] for clinical VQA. However, existing RFT frameworks optimize primarily for answer correctness through verifiable rewards [10], without explicitly addressing the alignment between model confidence and prediction accuracy. Unlike methods that elicit verbal confidence or simply reward high confidence, **CARE** uses confidence as a correctness-conditioned calibration signal within GRPO. Medical-CoT synthesis mainly provides a verified cold start that stabilizes reasoning format and answer extraction for subsequent confidence-aware RL.

To address this gap, we propose **CARE**—a **C**onfidence-**A**ware medical **RE**asoning framework that jointly optimizes diagnostic accuracy and confidence

calibration within a unified two-stage RFT pipeline. Our main contributions are as follows:

- **Confidence-Aware Reinforcement Fine-Tuning:** We integrate a novel Confidence-Aware Reward (**CAR**) into the GRPO framework, explicitly aligning the model's expressed confidence with diagnostic accuracy during RL optimization.
- **Scalable Medical-CoT Data Construction:** We design an automated synthesis pipeline to construct high-quality medical reasoning data for multiple clinical scenarios, providing structured diagnostic trajectories with verifiable conclusions.
- **Consistent Improvements in Accuracy and Calibration:** Evaluations across multiple Medical VQA benchmarks demonstrate that **CARE** achieves superior diagnostic accuracy while substantially reducing confidence miscalibration, establishing a reliable foundation for clinical decision support.

2 Method

2.1 Overview

As illustrated in Figure 2, the **CARE** framework consists of two phases: (1) *Medical-CoT Data Synthesis*, which constructs scalable, well-structured diagnostic reasoning paths from existing clinical datasets; and (2) *Two-Stage Optimization*, transitioning from SFT-based domain adaptation to GRPO reinforcement learning driven by a novel **Confidence-Aware Reward (CAR)** mechanism. This design jointly optimizes diagnostic accuracy and confidence calibration, mitigating the risk of confidence miscalibration in clinical judgments.

2.2 Scalable Medical-CoT Data Synthesis

To facilitate structured reasoning without relying solely on prohibitive expert annotations, we design an automated, reverse-thinking synthesis pipeline. Let $\mathcal{D}_{\text{VQA}} = \{(V_i, Q_i, Y_i)\}_{i=1}^{N}$ denote a standard Medical-VQA dataset comprising visual contexts V, text queries Q, and ground truth diagnoses Y. We employ a capable base MLLM, denoted as π_θ, to retroactively generate intermediate reasoning trajectories $\mathcal{T}$:

$$\mathcal{T}_i \sim \pi_\theta(V_i, Q_i, Y_i). \tag{1}$$

To ensure the clinical validity and logical coherence of $\mathcal{T}_i$, we enforce a strict structural template requiring explicit phase identifiers (e.g., visual analysis, differential diagnosis) and a conclusive summary. Crucially, rather than relying on raw generative outputs, we implement a rigorous verification mechanism. An auxiliary verifier (e.g., GPT-4o) evaluates each trajectory against the ground truth Y_i. A generated trajectory is admitted into the final training corpus $\mathcal{D}_{\text{CoT}}$ if and only if its reasoning chain logically deduces a conclusion that perfectly aligns with Y_i. This objective filtering acts as a scalable proxy for quality assurance, ensuring the model learns from goal-oriented, logically sound diagnostic paths.

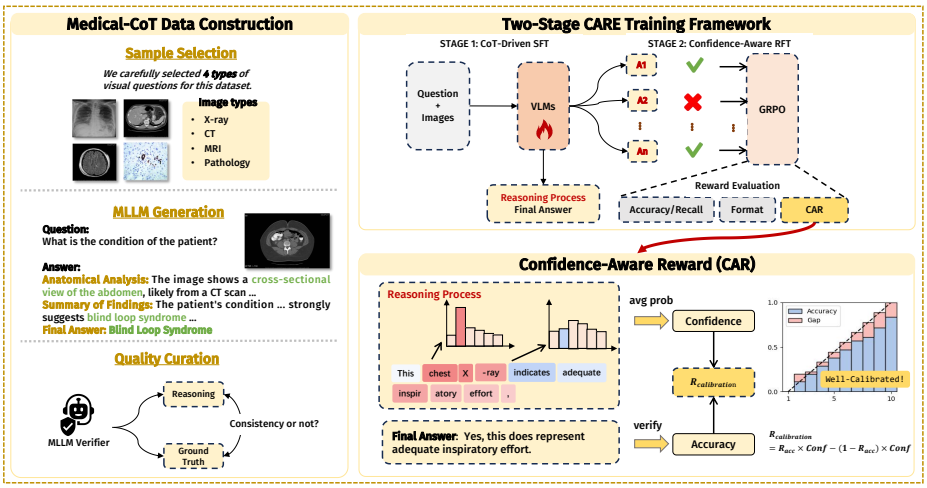

Fig. 2. *Overview of the **CARE** Framework.* The pipeline consists of two phases: *(1) Medical-CoT Data Construction* for synthesizing structured diagnostic reasoning paths, and *(2) Two-Stage Optimization*, comprising SFT cold-start and GRPO with the proposed **Confidence-Aware Reward (CAR)** mechanism to align confidence with diagnostic accuracy.

2.3 Two-Stage Optimization of CARE

Leveraging the curated $\mathcal{D}_{\text{CoT}}$, **CARE** is trained in two stages: SFT cold-start for domain adaptation, followed by confidence-aware GRPO-based reinforcement learning.

Phase I: SFT Cold Start. We conduct standard supervised fine-tuning on $\mathcal{D}_{\text{CoT}}$ to establish the model's capacity for structured clinical reasoning. For each tuple $(V, Q, \mathcal{T}, Y) \in \mathcal{D}_{\text{CoT}}$, the model maximizes the likelihood of the target sequence $S = [\mathcal{T}; Y]$:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(V,Q,S)\sim\mathcal{D}_{\text{CoT}}}\left[\sum_{t=1}^{|S|} \log \pi_\theta(s_t | V, Q, s_{<t})\right]. \tag{2}$$

This phase yields a reference policy π_{ref} that follows the structured reasoning format, serving as initialization for the RL phase.

Phase II: GRPO-based RL. We then optimize π_θ using GRPO, which eliminates the need for a separate value network. For each query (V, Q), the policy samples K candidate outputs $\{o_1, \ldots, o_K\}$. The objective maximizes group-relative advantages while constraining divergence from π_{ref}:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K} \min\left(\rho_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\right) A_i - \beta \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\right], \tag{3}$$

where $\rho_i = \frac{\pi_\theta(o_i|V,Q)}{\pi_\theta^{\text{old}}(o_i|V,Q)}$ is the importance weight, ϵ is the clip ratio, and β controls the KL penalty. The advantage A_i is computed by normalizing the composite reward R_i within the sampled group:

$$A_i = \frac{R_i - \mu(R_{1:K})}{\sigma(R_{1:K})}. \tag{4}$$

2.4 Confidence-Aware Reward (CAR) Formulation

Existing RFT frameworks rely on binary format or accuracy rewards, leaving confidence miscalibration unaddressed. We introduce the **Confidence-Aware Reward (CAR)** to close this gap by directly incorporating confidence alignment into the reward signal. The composite reward $R_i = R_{\text{form}} + R_{\text{out}} + R_{\text{calib}}$ consists of three components:

Format & Output Rewards ($R_{\text{form}}, R_{\text{out}}$). $R_{\text{form}} \in \{0, 1\}$ enforces the use of designated `<think>` and `<answer>` delimiters. R_{out} evaluates diagnostic correctness: for closed-ended tasks, it is an exact-match indicator $I(Y \subseteq a_i)$; for open-ended tasks, it uses a recall-based metric to assess the coverage of ground truth Y within the predicted answer a_i.

Calibration Reward (R_{calib}). This is the core component that distinguishes **CARE** from standard RFT. We use the model's answer-level predictive confidence as a calibration signal, rather than treating token probability as a complete epistemic uncertainty estimate. Specifically, for each output o_i, let $a_i = \{t_1, \ldots, t_{|a_i|}\}$ denote only the tokens inside the `<answer>` span, excluding reasoning tokens, formatting tokens, and special tokens. We compute:

$$C(a_i) = \frac{1}{|a_i|} \sum_{j=1}^{|a_i|} \pi_\theta(t_j \mid V, Q, t_{<j}). \tag{5}$$

The calibration reward ties this answer-level confidence to diagnostic correctness:

$$R_{\text{calib}}(o_i, Y) = R_{\text{out}} \cdot C(a_i) - \lambda(1 - R_{\text{out}}) \cdot C(a_i), \tag{6}$$

where λ controls the penalty on overconfident incorrect predictions. When the prediction is correct, higher answer confidence is rewarded; when it is incorrect, high confidence is penalized. This correctness-conditioned design differs from simply encouraging high confidence, and directly optimizes the confidence-accuracy alignment measured by ECE.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate CARE on three widely adopted Medical VQA benchmarks:

- **VQA-RAD** [11]: Focused on radiology, comprising 315 clinician-annotated images and 3,515 query-answer pairs.
- **SLAKE** [14]: A semantically-labeled, knowledge-enhanced dataset featuring 642 images and 14,000 bilingual pairs.
- **PathVQA** [7]: Dedicated to pathology, containing 4,998 images and 32,799 question-answer pairs.

Baselines. We evaluate CARE against state-of-the-art reasoning-focused Medical MLLMs, categorized by parameter scale: (1) *Compact-Scale Models ($\leq$ 3B):* Including Med-R1-3B [9] and MedVLM-R1-2B [15]. (2) *Standard-Scale Models (7B–8B):* Including Lingshu-7B [19], MedVLThinker-7B [8], Fleming-VL-8B [17], and MedMO-8B [4].

Evaluation Metrics. We assess model performance along three dimensions:

- *Diagnostic Accuracy:* We report Accuracy for closed-ended questions and Recall for open-ended questions, following standard Medical-VQA evaluation protocols.
- *Confidence Calibration:* We adopt Expected Calibration Error (ECE) [6] to measure the alignment between model confidence and actual accuracy. Confidence scores are partitioned into M equally spaced bins B_m, and ECE computes the weighted deviation:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|, \tag{7}$$

  where N is the total number of samples. Lower ECE indicates better calibration.
- *Hallucination Rate:* We employ Lingshu-32B [19] as a VLM judge to evaluate reasoning trajectories under a fixed prompt and rubric. The judge is run with deterministic decoding, and all methods are evaluated with identical inputs to ensure fairness. Each sample receives a normalized hallucination score r_i, and the overall HR is computed as:

$$\mathrm{HR} = \frac{C}{\mathbb{E}\left[(1 - r_i)^2 \, t_i\right]}, \tag{8}$$

  where t_i is the output token count and C is a global normalization constant ensuring $\mathrm{HR} \in [0, 1]$. The squared term $(1-r_i)^2$ applies a quadratic penalty that disproportionately suppresses samples with high hallucination rates, amplifying the contribution of heavily hallucinated outputs to the overall score.

3.2 Implementation Details

CARE is built upon Qwen2.5-VL-7B-Instruct [1]. The same model architecture serves as both the CoT data synthesizer in Section 2.2 and the policy model for subsequent training, ensuring consistency between the generated reasoning format and the optimization target. Training is conducted via full-parameter

fine-tuning on 6× NVIDIA A100 GPUs. During the *SFT Cold Start*, we use the AdamW optimizer with a learning rate of 1×10^{-5} and cosine annealing. In the *GRPO RL Phase*, we generate $K = 4$ rollouts per query with a batch size of 2 and `bfloat16` mixed precision. In the calibration reward of **CAR**, the penalty coefficient is set to $\lambda = 0.5$.

Table 1. Main results on Medical VQA benchmarks. The best and second-best results in each column are highlighted in **bold** and underlined, respectively. The same applies to Table 2 (**ACC** = accuracy, **ECE** = expected calibration error, **HR** = hallucination rate).

Model	VQA-RAD			SLAKE			PathVQA		
	ACC ↑	ECE ↓	HR ↓	ACC ↑	ECE ↓	HR ↓	ACC ↑	ECE ↓	HR ↓
Compact-Scale Models (≤ 3B)									
Med-R1-3B	0.513	0.451	0.195	0.596	0.369	0.257	0.387	0.584	0.155
MedVLM-R1-2B	0.532	0.448	0.200	0.504	0.474	0.237	0.387	0.587	0.168
Standard-Scale Models (7B–8B)									
MedMO-8B	0.647	0.264	0.365	0.816	0.277	0.368	0.563	0.383	0.223
Lingshu-7B	0.679	0.375	0.083	0.831	0.322	0.105	0.619	0.522	0.098
MedVLThinker-7B	0.637	0.413	0.071	0.678	0.424	0.099	0.652	0.569	0.082
Fleming-VL-8B	0.668	0.333	0.073	0.819	0.181	0.089	0.629	0.455	0.073
CARE-7B (Ours)	**0.767**	**0.202**	**0.048**	**0.873**	**0.115**	**0.070**	**0.689**	**0.290**	**0.059**

3.3 Experimental Results and In-depth Analysis

CARE simultaneously achieves the best accuracy, calibration, and lowest hallucination across all benchmarks. As shown in Table 1, existing medical reasoning models consistently exhibit trade-offs among these three dimensions: strong accuracy often comes with poor calibration or high hallucination, and vice versa. No baseline ranks first across all three metrics on any single benchmark. **CARE** is the only model to do so consistently, achieving the highest accuracy (e.g., 0.873 on SLAKE, 0.767 on VQA-RAD) while simultaneously obtaining the lowest ECE and HR. The substantial ECE reduction (e.g., 0.115 on SLAKE, a 36% relative improvement over the second-best Fleming-VL-8B) directly confirms the effectiveness of **CAR** in aligning confidence with accuracy. The consistently lowest HR (e.g., 0.048 on VQA-RAD) further validates our reverse-thinking data synthesis: answer-grounded reasoning paths reduce fabrication at the source by providing structured, goal-oriented CoT trajectories.
Different question types favor different training configurations. Table 2 reveals a clear pattern: for closed-ended questions, RL alone achieves the best accuracy and ECE across all three benchmarks (e.g., 0.864 ACC and 0.096 ECE on VQA-RAD), as the constrained answer space allows RL exploration

Table 2. Ablation study results for different training stages.

Training Stages	VQA-RAD				SLAKE				PathVQA			
	Open ↑	ECE ↓	Closed ↑	ECE ↓	Open ↑	ECE ↓	Closed ↑	ECE ↓	Open ↑	ECE ↓	Closed ↑	ECE ↓
Training-Free	0.483	0.485	0.658	0.341	0.513	0.456	0.690	0.287	0.150	0.823	0.664	0.326
SFT	0.520	0.468	0.629	0.359	0.803	0.188	0.767	0.225	0.349	0.640	0.857	0.133
RL	0.563	0.407	**0.864**	**0.096**	0.819	0.176	**0.861**	**0.119**	0.166	0.727	**0.955**	**0.020**
SFT+RL	**0.620**	**0.363**	0.658	0.323	**0.881**	**0.112**	0.779	0.209	**0.421**	**0.561**	0.863	0.121

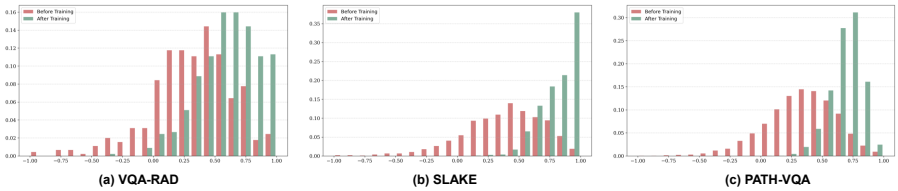

Fig. 3. *Confidence distributional shift.* Red and green bars represent the normalized confidence scores pre- and post-training, respectively. After training, the mean shifts rightward with increased magnitude, reflecting improved confidence-accuracy alignment.

to converge efficiently without SFT initialization. For open-ended questions, however, SFT+RL consistently dominates (e.g., 0.881 on SLAKE Open, 0.421 on PathVQA Open), since generating free-form diagnostic reasoning requires the structured CoT foundation established during SFT. Notably, SFT alone already yields substantial open-ended gains (e.g., SLAKE Open rises from 0.513 to 0.803), confirming that the CoT cold-start is the primary driver for open-ended performance, while RL further refines calibration on top of it. The results also reflect the complementary roles of the two stages: Medical-CoT provides a stable reasoning and answer-extraction format, especially for open-ended questions, while the RL stage further optimizes answer correctness and calibration through CAR.

CAR improves confidence-accuracy alignment rather than merely increasing confidence. Figure 3 compares the normalized confidence distributions before and after **CAR** alignment. The post-training distribution shifts toward higher-confidence regions, but this shift should be interpreted together with the simultaneous ECE reductions in Table 1. Since ECE directly measures the gap between confidence and empirical accuracy, the improved ECE indicates that the confidence shift is better aligned with diagnostic correctness rather than being a simple increase in likelihood. This supports the design of **CAR**, which rewards confident correct predictions while penalizing overconfident incorrect ones.

4 Conclusion

We present **CARE**, a framework that addresses confidence miscalibration in RFT-based medical reasoning models. By combining a scalable Medical-CoT synthesis pipeline with a Confidence-Aware Reward (**CAR**) mechanism within GRPO,

CARE explicitly aligns the model's expressed confidence with diagnostic accuracy during RL optimization. Experiments across three Medical VQA benchmarks show that **CARE** achieves the best diagnostic accuracy while simultaneously obtaining the lowest ECE and Hallucination Rate, demonstrating that accuracy and calibration can be jointly improved rather than traded off. These results suggest that incorporating confidence signals into the reward design is an effective path toward reliable medical AI systems. We sincerely hope this work encourages further exploration of calibration-aware training objectives in safety-critical domains beyond medical VQA.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 42394060 and 42394064, Ant Group Research Fund, and the Zhejiang University - Jolly Pharmaceutical Joint R&D Center for Intelligent Empowerment in Food and Medicine.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Bai, S., et al.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
2. 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 (2025), https://arxiv.org/abs/2501.17161
3. DeepSeek-AI, et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning (2025), https://arxiv.org/abs/2501.12948
4. Deria, A., Kumar, K., Dukre, A.M., Segal, E., Khan, S., Razzak, I.: Medmo: Grounding and understanding multimodal large language model for medical images (2026), https://arxiv.org/abs/2602.06965
5. Du, Y., Wang, Y., Kong, M., Liang, T., Long, Q., Chen, B., Zhu, Q.: Confidence calibration for multimodal llms: An empirical study through medical vqa. In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15965. Springer Nature Switzerland (September 2025)
6. Guo, C., Pleiss, G., Sun, Y., et al.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
7. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering (2020), https://arxiv.org/abs/2003.10286
8. Huang, X., Wu, J., Liu, H., Tang, X., Zhou, Y.: Medvlthinker: Simple baselines for multimodal medical reasoning (2025), https://arxiv.org/abs/2508.02669
9. Lai, Y., Zhong, J., Li, M., Zhao, S., Yang, X.: Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models (2025), https://arxiv.org/abs/2503.13939
10. Lambert, N., et al.: Tulu 3: Pushing frontiers in open language model post-training (2025), https://arxiv.org/abs/2411.15124
11. Lau, J., Gayen, S., Ben Abacha, A., et al.: A dataset of clinically generated visual questions and answers about radiology images. Scientific Data **5**(180251) (2018)
12. Li, C., Wong, C., Zhang, S., et al.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In: Advances in Neural Information Processing Systems. vol. 36 (2024)

13. Li, Z.Z., et al.: From system 1 to system 2: A survey of reasoning large language models (2025), https://arxiv.org/abs/2502.17419
14. Liu, B., Zhan, L.M., Xu, L., et al.: SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: IEEE 18th International Symposium on Biomedical Imaging. pp. 1650–1654 (2021)
15. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Cheng, O., Rueckert, D.: MedVLM-R1: Incentivizing medical reasoning capability of vision-language models (VLMs) via reinforcement learning. arXiv preprint (2025), https://arxiv.org/abs/2502.19634, arXiv:2502.19634
16. Shao, Z., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models (2024), https://arxiv.org/abs/2402.03300
17. Shu, Y., Liu, C., Chen, R., Li, D., Dai, B.: Fleming-vl: Towards universal medical visual reasoning with multimodal llms (2025), https://arxiv.org/abs/2511.00916
18. Sun, H., Jiang, Y., Lou, W., Zhang, Y., Li, W., Wang, L., Liu, M., Liu, L., Wang, X.: Chiron-o1: Igniting multimodal large language models towards generalizable medical reasoning via mentor-intern collaborative search (2025), https://arxiv.org/abs/2506.16962
19. Team, L., Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., Sun, Y., Shen, J., Wang, C., Tan, J., Zhao, D., Xu, T., Zhang, H., Rong, Y.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning (2025), https://arxiv.org/abs/2506.07044
20. Tian, K., Mitchell, E., Zhou, A., et al.: Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. arXiv preprint (2023), https://arxiv.org/abs/2305.14975, arXiv:2305.14975
21. Xiong, M., Hu, Z., Lu, X., et al.: Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. arXiv preprint (2023), https://arxiv.org/abs/2306.13063, arXiv:2306.13063
22. Xu, F., et al.: Towards large reasoning models: A survey of reinforced reasoning with large language models (2025), https://arxiv.org/abs/2501.09686