CARE: Causally-Aligned Reasoning Exploration for Medical Large Language Models

Yucheng Zhou[1,*], Peng Luo[1,*], Qianning Wang[2], Cheng-zhong Xu[1], and Jianbing Shen[1✉]

[1] SKL-IOTSC, CIS, University of Macau, Macau, China
[2] Auckland University of Technology, Auckland, New Zealand
yucheng.zhou@connect.um.edu.mo, jianbingshen@um.edu.mo

Abstract. Large Language Models (LLMs) have shown strong potential for medical reasoning, yet the scarcity and cost of expert-annotated data constrain their progress. While reinforcement learning offers a scalable alternative, standard outcome-based methods in medicine often suffer from autoregressive credit assignment failure and gradient variance explosion. This leads to the *"Right Answer, Wrong Reason"* trap, where models inadvertently reinforce spurious correlations and dataset shortcuts rather than valid clinical deduction. In this work, we propose **Causally-Aligned Reasoning Exploration (CARE)**, a theoretically grounded framework for intrinsic experience curation. CARE is built upon two rigorous conditions for high-quality training trajectories: Causal Sufficiency, which utilizes an agreement-based self-verification mechanism to mimic *do*-calculus interventions and effectively debias gradients; and Proximal Learnability, which employs dynamic entropy bounds to select experiences within the model's zone of proximal development for variance-bounded optimization. These rigorously filtered experiences are optimized via a dual-stream objective that combines on-policy group-relative exploration with difficulty-weighted experience replay. Extensive experiments on diverse medical multimodal and text-only benchmarks demonstrate that CARE consistently outperforms other strong competitors, substantially reducing correct-but-inconsistent reasoning and improving training stability.

Keywords: Medical Large Language Models · Reinforcement Learning · Causally-Aligned Reasoning

1 Introduction

Recent advances in Large Language Models (LLMs) have significantly expanded the scope of medical artificial intelligence, enabling AI systems to interpret

∗Equal contribution. ✉ Corresponding author: *Jianbing Shen*. This work was supported by the National Natural Science Foundation of China (No. 624B2002) and the Jiangyin Hi-tech Industrial Development Zone under the Taihu Innovation Scheme (EF2025-00003-SKL-IOTSC).

complex clinical data, reason over multimodal inputs, and provide diagnostic support across a wide range of medical tasks [25, 74]. The dominant training paradigm for such models remains Supervised Fine-Tuning (SFT) on curated medical datasets [22, 56]. While effective, SFT fundamentally depends on large volumes of expert-annotated data, which are costly, time-consuming, and often infeasible to scale in high-stakes medical domains. Consequently, there is growing interest in self-improving paradigms, particularly reinforcement learning (RL), where models optimize their reasoning strategies by learning from their own generated trajectories [12, 49, 70].

However, directly applying outcome-based RL (e.g., GRPO) to medical reasoning poses unique and critical challenges [69]. First, unlike domains such as mathematics or programming, where correctness can often be verified deterministically [59], medical reasoning lacks inexpensive and reliable automated verifiers. Relying solely on outcome supervision (e.g., ground-truth diagnoses) creates a severe misalignment between the reward signal and the actual reasoning process. In this work, we formally analyze this as the *"Right Answer, Wrong Reason"* trap: due to autoregressive credit assignment failure, outcome-based rewards inevitably reinforce spurious correlations and dataset shortcuts. The model learns to guess the correct label without engaging in valid clinical deduction. Such causally misaligned trajectories may artificially inflate benchmark accuracy but severely compromise the model's reliability and interpretability in real-world clinical settings.

Second, medical datasets exhibit extreme heterogeneity in information density, ranging from trivial recognition tasks to ambiguous, underspecified queries. Standard exploration strategies, which indiscriminately reinforce all correct rollouts, fail to account for the variance of the learning signal. As we demonstrate theoretically, unbounded sequence likelihoods destabilize the policy gradient. Trivial samples yield vanishing gradients, while highly uncertain or hallucinatory rationales introduce exploding variance that destroys optimization stability. Efficient self-improvement requires distinguishing *effective learning experiences* from statistical noise, ensuring optimization occurs strictly within a stable region of the model's competence.

To overcome these fundamental limitations, we argue that robust medical self-improvement requires shifting from simple outcome monitoring to rigorous intrinsic experience curation. We posit that a high-quality training trajectory must satisfy two theoretical conditions, which we rigorously formulate: (1) **Causal Sufficiency**: The generated rationale must be causally sufficient to recover the final answer. We prove that enforcing this condition acts as a *gradient debiasing operator*, ensuring the decision stems from logical reasoning rather than spurious priors; and (2) **Proximal Learnability**: The trajectory should fall within a dynamic, entropy-bounded window, the model's *zone of proximal development*. We mathematically guarantee that this constraint provides *variance-bounded optimization*, maximizing gradient utility while preventing instability.

Building on these theoretical foundations, we propose **Causally-Aligned Reasoning Exploration (CARE)**, a framework designed to seamlessly trans-

late our theoretical principles into a practical RL algorithm. CARE replaces naive filtration with a structured admission mechanism. To enforce causal sufficiency, we introduce an Agreement-Based Self-Verification protocol, which accepts an experience only if the model can reproduce the final decision when conditioned *solely* on the generated rationale (mimicking a *do*-calculus intervention). To control optimization variance, we implement a Learnability-Aware Exploration mechanism, dynamically selecting experiences with normalized sequence likelihoods that offer the highest information gain. These rigorously curated experiences are then optimized via a dual-stream objective, combining on-policy group-relative exploration with difficulty-weighted experience replay.

We evaluate CARE extensively on diverse medical multimodal and text-only benchmarks, including clinical VQA and complex diagnostic reasoning, and CARE consistently outperforms other models of a similar scale. Crucially, our analysis reveals that CARE not only improves accuracy but also significantly reduces correct-but-inconsistent reasoning, validating that enforcing causal alignment effectively mitigates shortcut learning.

Our contributions are summarized as follows:

- We theoretically prove that standard outcome-based RL in medicine suffers from autoregressive credit assignment failure and from an explosion of gradient variance. This formal analysis exposes the fundamental "Right Answer, Wrong Reason" trap that reinforces spurious correlations.
- We propose CARE, a framework that reformulates self-improvement as intrinsic curation governed by *Causal Sufficiency*. We implement this via Agreement-Based Self-Verification, which acts as a gradient debiasing operator to eliminate shortcut learning.
- We introduce a *Proximal Learnability* mechanism that filters trajectories based on dynamic entropy bounds. This theoretically guarantees variance-bounded optimization, ensuring the model learns solely from effective experiences within its zone of proximal development.
- Extensive experiments across multimodal and text-only medical benchmarks demonstrate that CARE outperforms other models of a similar scale.

2 Related Work

Foundational RL methods, evolving from Policy Gradient [53] and Actor-Critic [10, 34] to stable algorithms like PPO [47, 71] (addressing TRPO [46] instability) and sequence modeling via Decision Transformers [5, 28, 73], have fundamentally shaped LLMs. While early medical LLMs such as HealthGPT [31] and MMedLM [44] primarily utilized pre-training and instruction tuning [1], the adoption of RLHF [38] became standard for aligning models with medical values [64, 66]. To overcome context limitations and enhance clinical reasoning in complex scenarios [21, 62], researchers integrated multi-agent frameworks [15,43,57,58,72] and memory-augmented mechanisms [30,39]. Specialized reasoning methods have also emerged, including StarPO [61] and GRPO [9,50];

notably, GRPO eliminates the value model to stabilize updates and has been successfully applied to medical multimodal models [41,51,68] and mathematical reasoning [65]. Recent advancements focus on high-fidelity simulation [11, 57], and intermediate reasoning trajectories, exemplified by HuatuoGPT-o1 [3] (inspired by OpenAI o1 [20]). Furthermore, RL frameworks have expanded to multimodal domains [18] through models like MedCCO [45] and Med-R1 [26]. Current trends emphasize practical efficiency and comprehensive capabilities, e.g., multi-agent data generation [52], "RL + LoRA" efficient training [16, 42], and all-purpose processing (Lingshu [56] and Hulu-Med [22]).

3 Theoretical Analysis: Causal Misalignment and Optimization Stability

In this section, we provide a formal analysis of the optimization dynamics in self-improving LLMs. We first prove that standard outcome-based RL is structurally prone to reinforcing spurious correlations due to autoregressive credit assignment failure. We then demonstrate that our proposed **Causal Sufficiency** mechanism acts as a gradient debiasing operator, while **Proximal Learnability** guarantees variance-bounded optimization.

3.1 The Autoregressive Shortcut Trap

Let $x \in \mathcal{X}$ be a clinical query. The LLM generates a trajectory y composed of a latent rationale sequence $r = (r_1, \ldots, r_T)$ and a final answer a. The policy π_θ factorizes the joint probability via the chain rule:

$$\pi_\theta(r, a \mid x) = \underbrace{\pi_\theta(r \mid x)}_{\text{Rationale Generation}} \cdot \underbrace{\pi_\theta(a \mid x, r)}_{\text{Answer Prediction}}. \tag{1}$$

In standard outcome-based RL, the objective is to maximize the expected return $J(\theta) = \mathbb{E}_{y \sim \pi_\theta}[\mathcal{U}(a, a^*)]$, where $\mathcal{U}(a, a^*) = \mathbb{I}[a = a^*]$ is the binary correctness reward. The gradient estimator is:

$$\nabla_\theta J(\theta) \approx \mathbb{E}\left[\mathcal{U}(a, a^*)\left(\nabla_\theta \log \pi_\theta(r \mid x) + \nabla_\theta \log \pi_\theta(a \mid x, r)\right)\right]. \tag{2}$$

Definition 1 (Spurious Shortcut). *A rationale r_S is a* spurious shortcut *if it is logically insufficient to deduce a^* (i.e., $a^* \not\perp x \mid r_S$), but the model correctly predicts a^* due to residual correlations between x and a^* in the pre-trained weights.*

Theorem 1 (Credit Assignment Failure). *Under standard outcome supervision, if a spurious shortcut r_S leads to a correct answer a^*, the policy gradient strictly increases the likelihood of r_S:*

$$\mathbb{E}\left[\nabla_\theta \log \pi_\theta(r_S \mid x) \cdot \mathbb{I}[a = a^*]\right] > 0. \tag{3}$$

Proof. The scalar reward $\mathcal{U} = 1$ is broadcast to the entire sequence. Since $\nabla_\theta \log \pi_\theta(r_S \mid x)$ is additively coupled with the answer's gradient, the optimizer cannot distinguish whether the answer was derived *via* r_S or simply predicted *after* r_S. Consequently, valid reasoning and hallucinatory shortcuts are indiscriminately reinforced, cementing the "Right Answer, Wrong Reason" behavior.

3.2 Causal Sufficiency as Gradient Debiasing

To rectify this, we must ensure that the rationale r is a *sufficient statistic* for the answer a. We formalize this using the *do*-calculus intervention $do(R = r)$, which severs the dependence on the input x.

Definition 2 (Causal Sufficiency). *A trajectory (r, a) satisfies causal sufficiency if the answer a is recoverable from r alone under the current policy:*

$$\phi_{causal}(r, a) = \mathbb{I}\left[\arg\max_{\tilde{a}} \pi_{\bar{\theta}}(\tilde{a} \mid do(R = r)) = a\right], \tag{4}$$

where $\pi_{\bar{\theta}}$ represents the model in inference mode without access to x.

Theorem 2 (Gradient Debiasing). *Let $\mathcal{S}$ be the subspace of spurious shortcuts and $\mathcal{C}$ be the subspace of valid causal reasoning. Applying the filter ϕ_{causal} eliminates gradient contributions from $\mathcal{S}$:*

$$\forall r_S \in \mathcal{S}, \quad \nabla_\theta J_{CARE}(\theta)\big|_{r=r_S} \to 0. \tag{5}$$

Proof. For any $r_S \in \mathcal{S}$, the correctness of a^* relies on the confounder x. Under the intervention $do(R = r_S)$, x is masked. Since r_S lacks logical sufficiency, the conditional probability $\pi(a^* \mid r_S)$ drops significantly compared to $\pi(a^* \mid x, r_S)$, causing the verification check to fail ($\phi_{\text{causal}} = 0$). The gradient for this trajectory is zeroed out, effectively debiasing the update direction towards $\mathcal{C}$.

3.3 Proximal Learnability and Variance Control

Finally, we address the optimization stability. The high variance of medical query difficulty leads to unstable gradients. Let $\mathcal{L}_{\text{seq}}(y) = -\frac{1}{|y|} \log \pi_\theta(y|x)$ be the length-normalized negative log-likelihood (NLL).

Lemma 1 (Gradient Variance Bound). *The variance of the policy gradient estimator $\hat{g}$ is bounded by the second moment of the sequence NLL:*

$$Var(\hat{g}) \leq \mathcal{O}\left(\mathbb{E}\left[\mathcal{L}_{seq}(y)^2\right]\right). \tag{6}$$

Theorem 3 (Variance-Bounded Optimization). *By restricting training to a dynamic window $\mathcal{W} = [\tau_{low}, \tau_{high}]$ of NLLs (Proximal Learnability), we guarantee:*

1. ***Bounded Gradient Variance:*** *$Var(\hat{g}|y \in \mathcal{W}) \leq C \cdot \tau_{high}^2$, preventing explosion from hallucinations.*

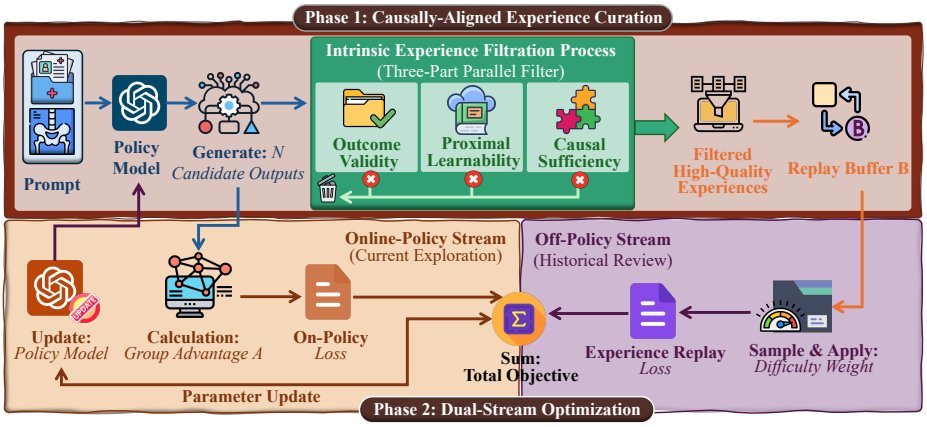

Fig. 1: Overview of the proposed CARE framework. The process consists of two phases: (1) **Causally-Aligned Experience Curation**, where candidate rollouts are evaluated through a three-part filter (Outcome Validity, Proximal Learnability, and Causal Sufficiency) to construct a high-fidelity learning signal; and (2) **Dual-Stream Optimization**, which robustly updates the policy using both on-policy exploration advantages and difficulty-weighted replay.

2. ***Non-Vanishing Signal:*** $||\mathbb{E}[\hat{g}|y \in \mathcal{W}]|| \geq \epsilon > 0$, *ensuring informative updates by rejecting trivial samples.*

Theorem 3 proves that our adaptive filtering does not merely curate data quality but mathematically constrains the optimization trajectory within a stable, low-variance region, enabling efficient self-improvement.

4 Methodology

We present Causally-Aligned Reasoning Exploration (CARE), a self-improvement framework designed to align policy optimization with valid clinical reasoning. As illustrated in Fig. 1, CARE translates the theoretical principles derived in Sec. 3 into a practical training algorithm. We shift the training paradigm from naive outcome supervision to intrinsic experience curation. We first detail our structured admission mechanism, which dynamically filters trajectories based on Proximal Learnability (variance control) and Causal Sufficiency (gradient debiasing). Finally, we describe the dual-stream optimization objective that leverages these curated experiences to stably improve the policy.

4.1 Problem Formulation

We consider a multimodal medical policy π_θ parameterized by θ, which processes a clinical query $x = (x_{\text{img}}, x_{\text{txt}})$ to generate a response y. To enable interpretability, we enforce a structured chain-of-thought format $y = (r, a)$, where r is the

generated rationale and a is the final diagnostic answer. The training set $\mathcal{D}$ contains query-answer pairs (x, a^*), but lacks expert-annotated rationales.

Our objective is to maximize the expected return $J(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta}[R(x, y)]$. As proved in Theorem 1, relying solely on a binary outcome reward $R(x, y) = \mathbb{I}[a = a^*]$ induces a gradient leak that reinforces spurious shortcuts. CARE addresses this by collecting N rollouts $\mathcal{Y}_x = \{y^{(i)}\}_{i=1}^N$ per prompt and rigorously curating them to construct an augmented reward signal and a high-fidelity replay buffer $\mathcal{B}$.

4.2 Causally-Aligned Experience Curation

To mitigate the reinforcement of spurious correlations and stabilize optimization, we introduce a unified admission function $\Phi(x, y) : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$. A trajectory is admitted if and only if it satisfies three hierarchical criteria.

Outcome Validity. The prerequisite for any valid experience is that the final decision aligns with the ground truth. Let $\mathcal{M}(\cdot)$ be a rule-based answer extraction function:

$$\phi_{\text{valid}}(x, y) = \mathbb{I}\left[\mathcal{M}(y) = a^*\right]. \tag{7}$$

Proximal Learnability (Variance Control). As theoretically established in Lemma 1, unbounded sequence negative log-likelihood (NLL) directly leads to gradient variance explosion. To ensure optimization strictly occurs within the model's *zone of proximal development* (Theorem 3), we quantify trajectory uncertainty using the length-normalized NLL:

$$\mathcal{L}_{\text{seq}}(y|x;\theta) = -\frac{1}{|y|}\sum_{t=1}^{|y|}\log \pi_\theta(y_t \mid x, y_{<t}). \tag{8}$$

We maintain a moving history buffer $\mathcal{H}$ of recent NLL values to dynamically compute a learnability window $[\tau_{\text{low}}, \tau_{\text{high}}]$ based on its quantiles. The learnability filter is:

$$\phi_{\text{learn}}(y) = \mathbb{I}\left[\tau_{\text{low}} \le \mathcal{L}_{\text{seq}}(y|x;\theta) \le \tau_{\text{high}}\right]. \tag{9}$$

This effectively rejects trivial shortcuts (vanishing gradients) and chaotic hallucinations (exploding variance), maximizing the stable Fisher information gain per update.

Causal Sufficiency via Agreement Verification. To explicitly debias the gradient (Theorem 2), we must ensure the decision a stems causally from the reasoning r. We empirically instantiate the $do(R = r)$ intervention using Agreement-Based Self-Verification. For a generated pair (r, a), we construct a verification prompt $\mathcal{T}(r)$ (e.g., *"Based solely on the reasoning: r, what is the diagnosis?"*) that masks

the original query x. We then query the model in inference mode ($\pi_{\bar{\theta}}$) to predict a new answer $\hat{a}$:

$$\hat{a} = \underset{\tilde{a}}{\operatorname{argmax}}\, \pi_{\bar{\theta}}(\tilde{a} \mid \mathcal{T}(r)). \tag{10}$$

The trajectory is causally sufficient only if the rationale is robust enough to independently reproduce the answer:

$$\phi_{\text{causal}}(r, a) = \mathbb{I}\left[\hat{a} = a\right]. \tag{11}$$

By penalizing trajectories where $\hat{a} \neq a$, this mechanism guarantees that r is a sufficient statistic for a, severing the dependence on dataset artifacts.

Unified Admission. The final CARE admission function serves as both the augmented RL reward and the gatekeeper for the replay buffer $\mathcal{B}$:

$$\Phi(x, y) = \phi_{\text{valid}}(x, y) \cdot \phi_{\text{learn}}(y) \cdot \phi_{\text{causal}}(r, a). \tag{12}$$

4.3 Dual-Stream Optimization

We optimize π_θ using a hybrid objective that combines on-policy Group Relative Policy Optimization (GRPO) with difficulty-weighted experience replay.

On-Policy Group Relative Exploration. For each prompt x, we generate a group of N candidate rollouts $\mathcal{Y}_x$. Instead of relying on a separate value network, we compute the advantage $A^{(i)}$ for the i-th rollout using the unified admission score $\Phi^{(i)} = \Phi(x, y^{(i)})$, standardized within the group:

$$A^{(i)} = \frac{\Phi^{(i)} - \mu(\Phi)}{\sigma(\Phi) + \epsilon}, \tag{13}$$

where $\mu(\Phi)$ and $\sigma(\Phi)$ are the mean and standard deviation of the admission scores in $\mathcal{Y}_x$. The on-policy objective is:

$$\mathcal{L}_{\text{on}}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \left[\frac{1}{N} \sum_{i=1}^{N} \left(A^{(i)} \log \pi_\theta(y^{(i)}|x) - \beta \mathbb{D}_{\text{KL}}[\pi_\theta || \pi_{\text{ref}}] \right) \right], \tag{14}$$

where the KL divergence penalty prevents excessive deviation from the reference policy π_{ref}.

Difficulty-Weighted Experience Replay. Simultaneously, we sample high-quality trajectories $(x, y) \sim \mathcal{B}$ that have successfully passed the curation pipeline. To further prioritize samples at the frontier of the model's capability, we apply an importance weight $w(y) \propto \mathcal{L}_{\text{seq}}(y|x; \theta)$ (normalized within the batch). The replay loss is:

$$\mathcal{L}_{\text{rep}}(\theta) = -\mathbb{E}_{(x,y) \sim \mathcal{B}} \left[w(y) \log \pi_\theta(y|x) \right]. \tag{15}$$

Algorithm 1 Causally-Aligned Reasoning Exploration (CARE)

Input: Dataset $\mathcal{D}$, Policy π_θ, Reference π_{ref}, Buffer $\mathcal{B} \leftarrow \emptyset$, History $\mathcal{H} \leftarrow \emptyset$.
while not converged **do**
  Sample a batch of prompts $\{x_j\}$ from $\mathcal{D}$.
  Update dynamic window $[\tau_{\text{low}}, \tau_{\text{high}}]$ using quantiles of $\mathcal{H}$.
  for each prompt x_j **do**
    Generate N rollouts $\{y_j^{(i)}\}_{i=1}^{N} \sim \pi_\theta(\cdot|x_j)$.
    for $i = 1$ **to** N **do**
      Compute NLL $\ell_j^{(i)} \leftarrow \mathcal{L}_{\text{seq}}(y_j^{(i)})$. Update $\mathcal{H}$.
      // Check Validity and Learnability
      if $\phi_{\text{valid}}(y_j^{(i)}) \wedge \phi_{\text{learn}}(y_j^{(i)})$ **then**
        // Check Causal Sufficiency via self-verification
        $\hat{a} \leftarrow \text{argmax}\, \pi_{\bar{\theta}}(a|\mathcal{T}(r_j^{(i)}))$.
        if $\hat{a} == a_j^{(i)}$ **then**
          $\Phi_j^{(i)} \leftarrow 1$, Add $(x_j, y_j^{(i)})$ to $\mathcal{B}$.
        else
          $\Phi_j^{(i)} \leftarrow 0$.
        end if
      else
        $\Phi_j^{(i)} \leftarrow 0$.
      end if
    end for
    Compute advantages $A_j^{(i)}$ using $\{\Phi_j^{(i)}\}_{i=1}^{N}$.
  end for
  Sample replay batch $\mathcal{B}_{\text{batch}} \sim \mathcal{B}$.
  Update θ by minimizing $\mathcal{L}_{\text{CARE}} = \mathcal{L}_{\text{on}} + \lambda \mathcal{L}_{\text{rep}}$.
end while

Total Objective. The final training objective seamlessly integrates exploration and stable exploitation:

$$\mathcal{L}_{\text{CARE}}(\theta) = \mathcal{L}_{\text{on}}(\theta) + \lambda \mathcal{L}_{\text{rep}}(\theta), \tag{16}$$

where λ controls the replay mixing ratio.

4.4 Training Algorithm

The CARE training procedure is summarized in Algorithm 1. The framework alternates between generating rollout groups, filtering them for causal sufficiency and learnability, and updating the policy via the dual-stream objective.

5 Experiments

5.1 Experimental Settings

Implementation Details. Our model, CARE, is built upon the HULU 7B medical vision-language backbone [22]. To isolate the effect of CARE from the

backbone architecture, we additionally report CARE (HuatuoGPT-V), which applies CARE to the HuatuoGPT-V 7B medical VLM [4]. Training is conducted on a mixture of medical datasets, including PMC-VQA [67], SLAKE [32], PathVQA [13], MedMCQA [40], and PubMedQA [24]. These datasets cover a broad spectrum of clinical imaging and professional knowledge reasoning tasks. We strictly ensure that there is no overlap between the training samples and the test sets of our evaluation benchmarks to prevent data contamination.

CARE is implemented using a rollout-based reinforcement learning framework on 8× A100 GPUs with a rollout batch size of 128 and an update batch size of 64. For each prompt, we generate $N = 8$ on-policy rollouts. The dynamic learnability window is governed by quantile thresholds $\alpha = 0.2$ and $\beta = 0.9$, calculated using a FIFO-based moving history buffer $\mathcal{H}$ of 2,000 NLL values. We set the replay loss weight $\lambda = 1.0$, the KL divergence penalty $\beta = 0.04$, and a fixed 50% replay ratio. The importance weight $w(y)$ is Softmax-normalized within each replay batch based on length-normalized NLLs. To ensure stable initialization, experience-based optimization and the replay buffer are activated only after the batch Pass@1 accuracy reaches 35%.

Benchmarks and Evaluation. We evaluate CARE across a comprehensive suite of clinical imaging and reasoning tasks, including multimodal benchmarks (OmniVQA [17], PMC-VQA [67], VQA-RAD [27], MMMU-Med [63], PathVQA [13], MedXQA [76], and SLAKE [32]) and professional text benchmarks (MMLU-Pro-Med [60], PubMedQA [24], MedMCQA [40], MedQA [23], and MMLU-Med [14]). To ensure reproducibility, all evaluation protocols, scoring scripts, and metrics strictly follow the official Hulu-Med GitHub repository.

5.2 Main Results

Medical Multimodal Benchmarks. Tab. 1 presents a comprehensive performance comparison across proprietary, general-purpose, and medical Vision-Language Models. CARE achieves the strongest overall performance among medical VLMs, consistently outperforming prior domain-specific models across all evaluated benchmarks. Notably, CARE significantly surpasses its backbone, Hulu-Med-7B, on all multimodal datasets. This gap demonstrates that our causally-aligned experience curation and dual-stream optimization provide substantial reasoning gains that go beyond mere architectural improvements. Furthermore, applying the CARE framework to HuatuoGPT-V leads to consistent performance boosts, confirming that CARE is backbone-agnostic and can effectively enhance the reasoning capabilities of existing medical VLMs. While large proprietary models remain competitive on certain benchmarks, CARE narrows the gap considerably.

Medical Text Benchmarks. We further evaluate the effectiveness of CARE on medical text-only reasoning benchmarks, as shown in Tab. 2. CARE consistently outperforms strong medical baselines, including HuatuoGPT-V and Hulu-Med, across all evaluated tasks. These results indicate that CARE enhances not only multimodal perception but also general medical knowledge deduction.

Table 1: Performance comparison on medical multimodal benchmarks across proprietary, general-purpose, and medical VLMs.

Models	OmniVQA	PMC-VQA	VQA-RAD	SLAKE	PathVQA	MedXQA	MMMU-Med
Proprietary Models							
GPT-4.1 [36]	75.5	55.2	65.0	72.2	55.5	45.2	75.2
GPT-4o [19]	67.5	49.7	61.0	71.2	55.5	44.3	62.8
Claude Sonnet 4 [54]	65.5	54.4	67.6	70.6	54.2	43.3	74.6
Gemini-2.5-Flash [55]	71.0	55.4	68.5	75.8	55.4	52.8	76.9
General-purpose MM							
Qwen2.5VL-7B [2]	63.6	51.9	63.2	66.8	44.1	20.1	50.6
Janus-Pro-7B [6]	59.6	50.1	49.7	55.2	35.4	18.4	36.1
InternVL2.5-8B [7]	81.3	51.3	59.4	69.0	42.1	21.7	53.5
InternVL3-8B [75]	79.1	53.8	65.4	72.8	48.6	22.4	59.2
Medical MM							
BiomedGPT [33]	27.9	27.6	16.6	13.6	11.3	-	24.9
Med-R1-2B [26]	-	47.4	39.0	54.5	15.3	21.1	34.8
MedVLM-R1-2B [41]	77.6	48.8	49.2	56.3	36.0	21.4	35.2
HealthGPT-M3 [31]	71.5	55.4	56.8	70.8	55.4	22.4	42.8
BioMedX2-8B [35]	66.0	41.8	55.7	54.1	34.6	21.9	39.8
LLaVA-Med-7B [29]	34.8	22.7	46.6	51.9	35.2	20.8	28.1
MedGemma-4B-IT [48]	70.7	49.2	72.3	78.2	48.1	25.4	43.2
HuatuoGPT-V 7B [4]	74.3	53.1	67.6	68.1	44.8	23.2	49.8
Lingshu-7B [56]	82.9	56.3	67.9	83.1	61.9	26.7	-
Hulu-Med-7B [22]	84.2	66.8	78.0	86.8	65.6	29.0	51.4
CARE-7B (HuatuoGPT-V)	75.8	54.6	69.0	69.8	46.3	25.4	51.2
CARE-7B	**85.6**	**68.2**	**79.3**	**88.1**	**67.1**	**31.2**	**53.0**

Specifically, the performance gains on complex benchmarks like MMLU-Med and MedQA suggest that enforcing causal sufficiency and proximal learnability effectively mitigates shortcut learning, thereby enhancing long-horizon reasoning and decision consistency in purely textual clinical settings.

5.3 Ablation Studies and Analysis

Methodological Baselines and Ablation Study. We first compare CARE against the standard RL baseline (Standard GRPO) and dissect our framework by systematically removing its core components. Tab. 3 presents the results on three representative benchmarks: PMC-VQA (multimodal), MedQA (textual), and MMMU-Med (complex reasoning). The Hulu-Med-7B model serves as our baseline. As shown in Tab. 3, applying standard outcome-based GRPO provides only marginal gains (+0.4% average) over the strong SFT baseline, suffering from noisy reward signals. Among our proposed components, removing Causal Sufficiency (ϕ_{causal}) results in the most noticeable performance drop compared to the full model, confirming that preventing shortcut learning is crucial for high-level reasoning. Removing Proximal Learnability (ϕ_{learn}) also degrades performance, indicating that unconstrained exploration on excessively hard or trivial samples introduces harmful gradient noise. Finally, the dual-stream difficulty replay is essential for ensuring stable exploitation, which further elevates performance.

Mitigating Shortcut Learning (Right Answer, Wrong Reason). To empirically validate that CARE mitigates the "Right Answer, Wrong Reason" trap, we con-

Table 2: Performance comparison on medical text benchmarks across proprietary, general-purpose, and medical models.

Models	MMLU-Pro-Med	MedXQA	PubMedQA	MedMCQA	MedQA	MMLU-Med
Proprietary Models						
GPT-4.1 [36]	78.0	30.9	75.6	77.7	89.1	89.6
o3-mini [37]	78.1	35.4	73.6	60.6	74.5	87.0
GPT-4o [19]	75.6	25.9	71.8	76.9	89.2	88.2
Claude Sonnet 4 [54]	79.5	33.6	78.6	79.3	92.1	91.3
Gemini-2.5-Flash [55]	70.0	35.6	73.8	73.6	91.2	84.2
Deepseek-V3 [8]	74.6	20.0	77.7	88.0	51.0	86.5
General-purpose MM						
Qwen2.5VL-7B [2]	50.5	12.8	76.4	52.6	57.3	73.4
Janus-Pro-7B [6]	20.2	10.0	72.0	37.5	37.4	46.4
InternVL2.5-8B [7]	50.6	11.6	76.4	52.4	53.7	74.2
InternVL3-8B [75]	57.9	13.1	75.4	57.7	62.1	77.5
Medical MM						
MedVLM-R1-2B [41]	24.9	11.8	66.4	39.7	42.3	51.8
BioMedX2-8B [35]	40.8	13.4	75.2	52.9	58.9	68.6
MedGemma-4B-IT [48]	38.6	12.8	72.2	52.2	56.2	66.7
HealthGPT-M3 [31]	38.3	11.5	57.8	54.2	55.0	72.5
LLaVA-Med-7B [29]	16.6	9.9	26.4	39.4	42.0	50.6
HuatuoGPT-V 7B [4]	44.6	10.1	72.8	51.2	52.9	69.3
Lingshu-7B [56]	50.4	16.5	76.6	55.9	63.3	74.5
Hulu-Med-7B [22]	60.6	19.6	77.4	67.6	73.5	79.5
CARE-7B (HuatuoGPT-V)	46.3	12.2	73.9	53.0	54.6	71.0
CARE-7B	**62.4**	**22.1**	**78.6**	**69.1**	**75.0**	**81.1**

ducted a clinical consistency evaluation. We randomly sampled 500 queries from the MedQA test set where *both* Standard GRPO and CARE predicted the correct final answer. The generated rationales were then evaluated by a panel of three human experts through majority voting, classifying them into: (1) Valid Causal Reasoning, (2) Spurious Shortcut (correct answer but logically insufficient or relying on statistical priors), and (3) Unrelated Hallucination. As shown in Fig. 2(Left), although Standard GRPO yields correct outcomes on these samples, **45.0%** of its rationales rely on spurious correlations. In contrast, CARE increases the proportion of Valid Causal Reasoning to **86.2%**. This corroborates Theorem 2, demonstrating that our self-verification mechanism successfully severs the model's reliance on dataset artifacts in favor of causally-aligned deduction.

Optimization Stability via Proximal Learnability. In Sec. 3, we proved that bounding the sequence NLL stabilizes gradient variance (Theorem 3). Fig. 2(Middle) visualizes this by plotting the Pass@1 evaluation accuracy on a held-out validation batch. Standard GRPO exhibits significant instability, marked by sharp spikes and reward degradation, as it indiscriminately reinforces low-probability hallucinatory trajectories that coincidentally lead to correct answers. In contrast, CARE's Proximal Learnability filter (ϕ_{learn}) rejects both trivial (vanishing gradient) and chaotic (exploding variance) samples. This constrains the policy to a stable region, ensuring monotonic improvement and a higher convergence ceiling.

Table 3: Ablation Study and Baseline Comparison. Evaluation of different training paradigms starting from the Hulu-Med-7B base model.

Method	PMC-VQA	MedQA	MMMU-Med	Avg.
Methodological Baselines				
Hulu-Med-7B (Base / SFT)	66.8	73.5	51.4	63.9
+ Standard GRPO (Outcome-only)	67.2	73.8	51.8	64.3
CARE Ablations (Starting from Base)				
CARE w/o Causal Sufficiency (ϕ_{causal})	67.5	74.2	52.1	64.6
CARE w/o Proximal Learnability (ϕ_{learn})	67.8	74.4	52.4	64.9
CARE w/o Difficulty Replay ($\lambda = 0$)	68.0	74.7	52.6	65.1
CARE (Ours)	**68.2**	**75.0**	**53.0**	**65.4**

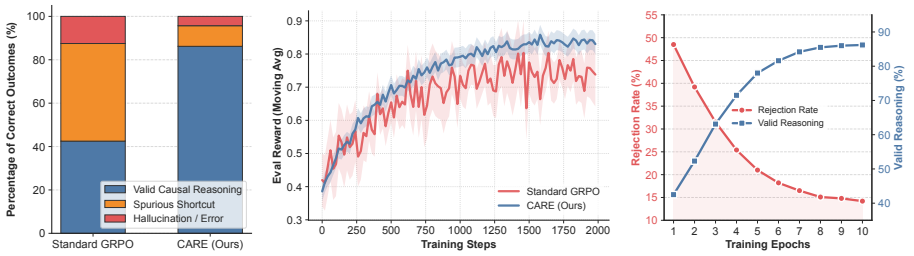

Fig. 2: (Left) Reasoning Consistency Analysis: Distribution of reasoning validity for samples where the final answer was predicted correctly. **(Middle)** Training Dynamics of Standard GRPO and CARE. **(Right)** Dynamics of Causal Alignment: Self-Verification Rejection Rate (red circles) and Valid Causal Reasoning (blue squares).

Dynamics of Causal Self-Verification. We evaluate our agreement-based self-verification mechanism (ϕ_{causal}) to ensure it is not excessively restrictive. Fig. 2 (Right) tracks the rejection rate across training epochs. Initially, 48.5% of correct trajectories are rejected, confirming that early-stage models heavily rely on superficial shortcuts rather than logical sufficiency. As CARE reinforces causally aligned trajectories, this rejection rate steadily declines to 15%, while valid causal reasoning simultaneously climbs to 86.2% (Fig. 2(Right), right axis). This inverse correlation shows that the model successfully internalizes structural causal constraints, shifting its behavior from dataset guessing to robust clinical deduction.

Empirical Validation of Gradient Variance (Theorem 3). We empirically validate Theorem 3 by tracking policy gradient L_2 norms during training. As shown in Fig. 3(Left), Standard GRPO suffers from frequent spikes due to indiscriminately updating on "chaotic" hallucinatory trajectories with unbounded NLLs that coincidentally yield correct answers. Conversely, CARE's Proximal Learnability filter (ϕ_{learn}) effectively rejects these variance-exploding samples. The resulting gradient L_2 norm remains tightly bounded and smooth, providing empirical proof for Theorem 3 and explaining the superior training stability observed in our experiments.

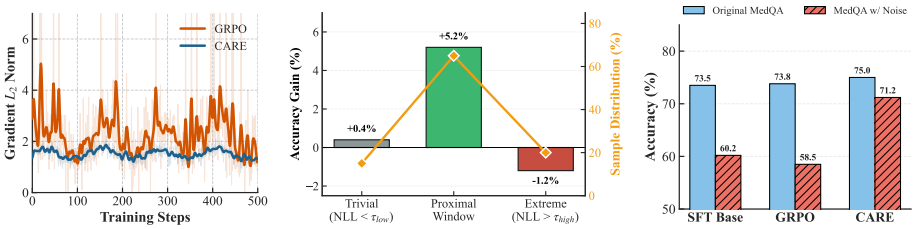

Fig. 3: (Left) Gradient L_2 Norm Tracking of CARE and GRPO. **(Middle)** Efficacy of the Proximal Window. **(Right)** Robustness to Spurious Clinical Distractors.

Efficacy of the Proximal Development Zone. To further dissect *why* the dynamic learnability window is crucial, we isolated the generated rollouts into three distinct NLL regions: Trivial (NLL $< \tau_{\text{low}}$), Proximal Window ($\tau_{\text{low}} \leq$ NLL $\leq \tau_{\text{high}}$), and Extreme Hard (NLL $> \tau_{\text{high}}$). We then conducted separate isolated training sessions utilizing only the experiences from each specific region. As shown in Fig. 3(Middle), relying solely on *Trivial* samples yields a negligible accuracy gain (+0.4%), as these samples provide vanishing gradients and minimal information gain. Conversely, forcing the model to learn from *Extreme Hard* samples results in a performance degradation (-1.2%) due to the introduction of chaotic, noisy signals that corrupt learned weights. The vast majority of the performance improvement (+5.2%) is derived entirely from the *Proximal Window*. This analysis validates our dynamic filtering mechanism: efficient self-improvement requires isolating the effective learning experiences from statistical noise.

Causal Sufficiency vs. Spurious Shortcuts (Theorem 2). Theorem 2 posits that our agreement-based self-verification acts as a *do*-calculus intervention to debias the gradient from spurious shortcuts. To validate that agreement-based self-verification acts as a *do*-calculus intervention, we tested OOD robustness by injecting irrelevant symptoms and noise into MedQA queries. As shown in Fig. 3(Right), the SFT Base model experiences a 13.3% performance drop under perturbation. Strikingly, Standard GRPO suffers an even more severe drop of 15.3%. This perfectly demonstrates the "Autoregressive Shortcut Trap" (Theorem 1): unconstrained outcome-based RL inadvertently reinforces the model's reliance on superficial keyword matching, making it highly fragile. In stark contrast, CARE exhibits remarkable robustness, degrading by only 3.8%. This confirms that enforcing decision-rationale consistency (ϕ_{causal}) successfully severs the model's dependence on dataset artifacts, forcing it to extract the true causal chain required to reach the diagnosis.

6 Conclusion

In this work, we introduced Causally-Aligned Reasoning Exploration, a self-improvement framework designed to mitigate the *causal misalignment* inherent

in outcome-based medical reinforcement learning. By shifting the optimization paradigm from naive outcome monitoring to rigorous intrinsic experience curation, governed by causal sufficiency and proximal learnability, CARE ensures that policy updates are driven by valid clinical deduction rather than spurious dataset shortcuts. Extensive evaluations across diverse multimodal and text-only benchmarks demonstrate that CARE consistently outperforms strong SFT and GRPO baselines while significantly enhancing reasoning consistency.

References

1. Alqahtani, M.Q., Albarakati, A., Alotaibi, F.: Refining medical large language models: key insights from instruction tuning. PeerJ Comput. Sci. **11**, e3216 (2025). https://doi.org/10.7717/PEERJ-CS.3216, https://doi.org/10.7717/peerj-cs.3216
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. CoRR **abs/2502.13923** (2025). https://doi.org/10.48550/ARXIV.2502.13923, https://doi.org/10.48550/arXiv.2502.13923
3. Chen, J., Cai, Z., Ji, K., Wang, X., Liu, W., Wang, R., Hou, J., Wang, B.: Huatuogpt-o1, towards medical complex reasoning with llms. CoRR **abs/2412.18925** (2024). https://doi.org/10.48550/ARXIV.2412.18925, https://doi.org/10.48550/arXiv.2412.18925
4. 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. CoRR **abs/2406.19280** (2024). https://doi.org/10.48550/ARXIV.2406.19280, https://doi.org/10.48550/arXiv.2406.19280
5. Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., Mordatch, I.: Decision transformer: Reinforcement learning via sequence modeling. In: Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 15084–15097 (2021), https://proceedings.neurips.cc/paper/2021/hash/7f489f642a0ddb10272b5c31057f0663-Abstract.html
6. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. CoRR **abs/2501.17811** (2025). https://doi.org/10.48550/ARXIV.2501.17811, https://doi.org/10.48550/arXiv.2501.17811
7. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., Gu, L., Wang, X., Li, Q., Ren, Y., Chen, Z., Luo, J., Wang, J., Jiang, T., Wang, B., He, C., Shi, B., Zhang, X., Lv, H., Wang, Y., Shao, W., Chu, P., Tu, Z., He, T., Wu, Z., Deng, H., Ge, J., Chen, K., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. CoRR **abs/2412.05271** (2024). https://doi.org/10.48550/ARXIV.2412.05271, https://doi.org/10.48550/arXiv.2412.05271

8. DeepSeek-AI: Deepseek-v3 technical report. CoRR **abs/2412.19437** (2024). https://doi.org/10.48550/ARXIV.2412.19437, https://doi.org/10.48550/arXiv.2412.19437
9. DeepSeek-AI: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. CoRR **abs/2501.12948** (2025). https://doi.org/10.48550/ARXIV.2501.12948, https://doi.org/10.48550/arXiv.2501.12948
10. Degris, T., White, M., Sutton, R.S.: Off-policy actor-critic. CoRR **abs/1205.4839** (2012), http://arxiv.org/abs/1205.4839
11. Dou, C., Liu, C., Yang, F., Li, F., Jia, J., Chen, M., Ju, Q., Wang, S., Dang, S., Li, T., Zeng, X., Zhou, Y., Zhu, C., Pan, D., Deng, F., Ai, G., Dong, G., Zhang, H., Tai, J., Hong, J., Lu, K., Sun, L., Guo, P., Ma, Q., Xin, R., Yang, S., Zhang, S., Mo, Y., Liang, Z., Zhang, Z., Cui, H., Zhu, Z., Wang, X.: Baichuan-m2: Scaling medical capability with large verifier system. CoRR **abs/2509.02208** (2025). https://doi.org/10.48550/ARXIV.2509.02208, https://doi.org/10.48550/arXiv.2509.02208
12. Gülçehre, Ç., Paine, T.L., Srinivasan, S., Konyushkova, K., Weerts, L., Sharma, A., Siddhant, A., Ahern, A., Wang, M., Gu, C., Macherey, W., Doucet, A., Firat, O., de Freitas, N.: Reinforced self-training (rest) for language modeling. CoRR **abs/2308.08998** (2023). https://doi.org/10.48550/ARXIV.2308.08998, https://doi.org/10.48550/arXiv.2308.08998
13. He, X., Zhang, Y., Mou, L., Xing, E.P., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. CoRR **abs/2003.10286** (2020), https://arxiv.org/abs/2003.10286
14. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. CoRR **abs/2009.03300** (2020), https://arxiv.org/abs/2009.03300
15. Hong, S., Zheng, X., Chen, J., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S.K.S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C.: Metagpt: Meta programming for multi-agent collaborative framework. CoRR **abs/2308.00352** (2023). https://doi.org/10.48550/ARXIV.2308.00352, https://doi.org/10.48550/arXiv.2308.00352
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: Lora: Low-rank adaptation of large language models. CoRR **abs/2106.09685** (2021), https://arxiv.org/abs/2106.09685
17. Hu, Y., Li, T., Lu, Q., Shao, W., He, J., Qiao, Y., Luo, P.: Omnimed-vqa: A new large-scale comprehensive evaluation benchmark for medical LVLM. CoRR **abs/2402.09181** (2024). https://doi.org/10.48550/ARXIV.2402.09181, https://doi.org/10.48550/arXiv.2402.09181
18. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. CoRR **abs/2503.06749** (2025). https://doi.org/10.48550/ARXIV.2503.06749, https://doi.org/10.48550/arXiv.2503.06749
19. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Madry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A.T., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kondrich, A., Tulloch, A., Mishchenko, A., Baek, A., Jiang, A., Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pantuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky, B., Wang, B., Zweig, B., Hoover, B., Samic, B.,

McGrew, B., Spero, B., Giertler, B., Cheng, B., Lightcap, B., Walkin, B., Quinn, B., Guarraci, B., Hsu, B., Kellogg, B., Eastman, B., Lugaresi, C., Wainwright, C.L., Bassin, C., Hudson, C., Chu, C., Nelson, C., Li, C., Shern, C.J., Conger, C., Barette, C., Voss, C., Ding, C., Lu, C., Zhang, C., Beaumont, C., Hallacy, C., Koch, C., Gibson, C., Kim, C., Choi, C., McLeavey, C., Hesse, C., Fischer, C., Winter, C., Czarnecki, C., Jarvis, C., Wei, C., Koumouzelis, C., Sherburn, D.: Gpt-4o system card. CoRR **abs/2410.21276** (2024). https://doi.org/10.48550/ARXIV.2410.21276, https://doi.org/10.48550/arXiv.2410.21276

20. Jaech, A., Kalai, A., Lerer, A., Richardson, A., El-Kishky, A., Low, A., Helyar, A., Madry, A., Beutel, A., Carney, A., Iftimie, A., Karpenko, A., Passos, A.T., Neitz, A., Prokofiev, A., Wei, A., Tam, A., Bennett, A., Kumar, A., Saraiva, A., Vallone, A., Duberstein, A., Kondrich, A., Mishchenko, A., Applebaum, A., Jiang, A., Nair, A., Zoph, B., Ghorbani, B., Rossen, B., Sokolowsky, B., Barak, B., McGrew, B., Minaiev, B., Hao, B., Baker, B., Houghton, B., McKinzie, B., Eastman, B., Lugaresi, C., Bassin, C., Hudson, C., Li, C.M., de Bourcy, C., Voss, C., Shen, C., Zhang, C., Koch, C., Orsinger, C., Hesse, C., Fischer, C., Chan, C., Roberts, D., Kappler, D., Levy, D., Selsam, D., Dohan, D., Farhi, D., Mely, D., Robinson, D., Tsipras, D., Li, D., Oprica, D., Freeman, E., Zhang, E., Wong, E., Proehl, E., Cheung, E., Mitchell, E., Wallace, E., Ritter, E., Mays, E., Wang, F., Such, F.P., Raso, F., Leoni, F., Tsimpourlas, F., Song, F., von Lohmann, F., Sulit, F., Salmon, G., Parascandolo, G., Chabot, G., Zhao, G., Brockman, G., Leclerc, G., Salman, H., Bao, H., Sheng, H., Andrin, H., Bagherinezhad, H., Ren, H., Lightman, H., Chung, H.W., Kivlichan, I., O'Connell, I., Osband, I., Gilaberte, I.C., Akkaya, I.: Openai o1 system card. CoRR **abs/2412.16720** (2024). https://doi.org/10.48550/ARXIV.2412.16720, https://doi.org/10.48550/arXiv.2412.16720
21. Jiang, S., Wang, Y., Chen, R., Zhang, Y., Luo, R., Lei, B., Song, S., Feng, Y., Sun, J., Wu, J., Liu, Z.: CAPO: reinforcing consistent reasoning in medical decision-making. CoRR **abs/2506.12849** (2025). https://doi.org/10.48550/ARXIV.2506.12849, https://doi.org/10.48550/arXiv.2506.12849
22. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., Hao, J., Chen, Z., Wu, R., Tang, T., Lv, J., Xu, H., Wang, H., Xiao, J., Feng, B., Zhu, F., Li, K., Xie, W., Sun, J., Wu, J., Liu, Z.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. CoRR **abs/2510.08668** (2025). https://doi.org/10.48550/ARXIV.2510.08668, https://doi.org/10.48550/arXiv.2510.08668
23. Jin, D., Pan, E., Oufattole, N., Weng, W., Fang, H., Szolovits, P.: What disease does this patient have? A large-scale open domain question answering dataset from medical exams. CoRR **abs/2009.13081** (2020), https://arxiv.org/abs/2009.13081
24. Jin, Q., Dhingra, B., Liu, Z., Cohen, W.W., Lu, X.: Pubmedqa: A dataset for biomedical research question answering. CoRR **abs/1909.06146** (2019), http://arxiv.org/abs/1909.06146
25. Kan, Z., Gan, W., Qi, Z., Yu, P.S.: Advances in large language models for medicine. CoRR **abs/2509.18690** (2025). https://doi.org/10.48550/ARXIV.2509.18690, https://doi.org/10.48550/arXiv.2509.18690
26. Lai, Y., Zhong, J., Li, M., Zhao, S., Yang, X.: Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. CoRR **abs/2503.13939** (2025). https://doi.org/10.48550/ARXIV.2503.13939, https://doi.org/10.48550/arXiv.2503.13939

27. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)
28. Levine, S., Kumar, A., Tucker, G., Fu, J.: Offline reinforcement learning: Tutorial, review, and perspectives on open problems. CoRR **abs/2005.01643** (2020), https://arxiv.org/abs/2005.01643
29. 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. CoRR **abs/2306.00890** (2023). https://doi.org/10.48550/ARXIV.2306.00890, https://doi.org/10.48550/arXiv.2306.00890
30. Li, G., Hammoud, H.A.A.K., Itani, H., Khizbullin, D., Ghanem, B.: CAMEL: communicative agents for "mind" exploration of large scale language model society. CoRR **abs/2303.17760** (2023). https://doi.org/10.48550/ARXIV.2303.17760, https://doi.org/10.48550/arXiv.2303.17760
31. Lin, T., Zhang, W., Li, S., Yuan, Y., Yu, B., Li, H., He, W., Jiang, H., Li, M., Song, X., Tang, S., Xiao, J., Lin, H., Zhuang, Y., Ooi, B.C.: Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. CoRR **abs/2502.09838** (2025). https://doi.org/10.48550/ARXIV.2502.09838, https://doi.org/10.48550/arXiv.2502.09838
32. Liu, B., Zhan, L., Xu, L., Ma, L., Yang, Y., Wu, X.: SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. CoRR **abs/2102.09542** (2021), https://arxiv.org/abs/2102.09542
33. Luo, Y., Zhang, J., Fan, S., Yang, K., Wu, Y., Qiao, M., Nie, Z.: Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. CoRR **abs/2308.09442** (2023). https://doi.org/10.48550/ARXIV.2308.09442, https://doi.org/10.48550/arXiv.2308.09442
34. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M.A.: Playing atari with deep reinforcement learning. CoRR **abs/1312.5602** (2013), http://arxiv.org/abs/1312.5602
35. Mullappilly, S.S., Kurpath, M.I., Pieri, S., Alseiari, S.Y., Cholakkal, S., Aldahmani, K., Khan, F., Anwer, R.M., Khan, S., Baldwin, T., Cholakkal, H.: Bimedix2: Bio-medical expert LMM for diverse medical modalities. CoRR **abs/2412.07769** (2024). https://doi.org/10.48550/ARXIV.2412.07769, https://doi.org/10.48550/arXiv.2412.07769
36. OpenAI, A.K., Yu, J., Hallman, J., et al.: Introducing gpt-4.1 in the api. https://openai. com/index/gpt-4-1/ (2025)
37. OpenAI, T.: Introducing openai o3 and o4-mini. https://openai. com/index/introducing-o3-and-o4-mini/ (2025)
38. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P.F., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html
39. Packer, C., Fang, V., Patil, S.G., Lin, K., Wooders, S., Gonzalez, J.E.: Memgpt: Towards llms as operating systems. CoRR **abs/2310.08560** (2023). https://doi.org/10.48550/ARXIV.2310.08560, https://doi.org/10.48550/arXiv.2310.08560

40. Pal, A., Umapathi, L.K., Sankarasubbu, M.: Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. CoRR **abs/2203.14371** (2022). https://doi.org/10.48550/ARXIV.2203.14371, https://doi.org/10.48550/arXiv.2203.14371
41. 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. CoRR **abs/2502.19634** (2025). https://doi.org/10.48550/ARXIV.2502.19634, https://doi.org/10.48550/arXiv.2502.19634
42. Pham, T., Ngo, C.: RARL: improving medical VLM reasoning and generalization with reinforcement learning and lora under data and hardware constraints. CoRR **abs/2506.06600** (2025). https://doi.org/10.48550/ARXIV.2506.06600, https://doi.org/10.48550/arXiv.2506.06600
43. Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., Yang, C., Chen, W., Su, Y., Cong, X., Xu, J., Li, D., Liu, Z., Sun, M.: Chatdev: Communicative agents for software development. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024. pp. 15174–15186. Association for Computational Linguistics (2024). https://doi.org/10.18653/V1/2024.ACL-LONG.810, https://doi.org/10.18653/v1/2024.acl-long.810
44. Qiu, P., Wu, C., Zhang, X., Lin, W., Wang, H., Zhang, Y., Wang, Y., Xie, W.: Towards building multilingual language model for medicine. CoRR **abs/2402.13963** (2024). https://doi.org/10.48550/ARXIV.2402.13963, https://doi.org/10.48550/arXiv.2402.13963
45. Rui, S., Chen, K., Ma, W., Wang, X.: Improving medical reasoning with curriculum-aware reinforcement learning. CoRR **abs/2505.19213** (2025). https://doi.org/10.48550/ARXIV.2505.19213, https://doi.org/10.48550/arXiv.2505.19213
46. Schulman, J., Levine, S., Moritz, P., Jordan, M.I., Abbeel, P.: Trust region policy optimization. CoRR **abs/1502.05477** (2015), http://arxiv.org/abs/1502.05477
47. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. CoRR **abs/1707.06347** (2017), http://arxiv.org/abs/1707.06347
48. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A.P., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., Chen, J., Mahvar, F., Yatziv, L., Chen, T.L., Sterling, B., Baby, S.A., Baby, S.M., Lai, J., Schmidgall, S., Yang, L., Chen, K., Bjornsson, P., Reddy, S., Brush, R., Philbrick, K., Asiedu, M., Mezerreg, I., Hu, H., Yang, H., Tiwari, R., Jansen, S., Singh, P., Liu, Y., Azizi, S., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J., Ramos, S., Yvinec, E., Casbon, M., Buchatskaya, E., Alayrac, J., Lepikhin, D., Feinberg, V., Borgeaud, S., Andreev, A., Hardin, C., Dadashi, R., Hussenot, L., Joulin, A., Bachem, O., Matias, Y., Chou, K., Hassidim, A., Goel, K., Farabet, C., Barral, J.K., Warkentin, T., Shlens, J., Fleet, D.J., Cotruta, V., Sanseviero, O., Martins, G., Kirk, P., Rao, A., Shetty, S., Steiner, D.F., Kirmizibayrak, C., Pilgrim, R., Golden, D., Yang, L.: Medgemma technical report. CoRR **abs/2507.05201** (2025). https://doi.org/10.48550/ARXIV.2507.05201, https://doi.org/10.48550/arXiv.2507.05201

49. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y.K., Wu, Y., Guo, D.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. CoRR **abs/2402.03300** (2024). https://doi.org/10.48550/ARXIV.2402.03300, https://doi.org/10.48550/arXiv.2402.03300
50. Song, H., Zhou, Y., Shen, J., Cheng, Y.: From broad exploration to stable synthesis: Entropy-guided optimization for autoregressive image generation. In: The Fourteenth International Conference on Learning Representations (2026)
51. Su, Y., Li, T., Liu, J., Ma, C., Ning, J., Tang, C., Ju, S., Ye, J., Chen, P., Hu, M., Tang, S., Liu, L., Fu, B., Shao, W., Hu, X., Liao, X., Ji, Y., He, J.: GMAI-VL-R1: harnessing reinforcement learning for multimodal medical reasoning. CoRR **abs/2504.01886** (2025). https://doi.org/10.48550/ARXIV.2504.01886, https://doi.org/10.48550/arXiv.2504.01886
52. Sun, Y., Qian, X., Xu, W., Zhang, H., Xiao, C., Li, L., Rong, Y., Huang, W., Bai, Q., Xu, T.: Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. CoRR **abs/2506.09513** (2025). https://doi.org/10.48550/ARXIV.2506.09513, https://doi.org/10.48550/arXiv.2506.09513
53. Sutton, R.S., McAllester, D.A., Singh, S., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. In: Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]. pp. 1057–1063. The MIT Press (1999), http://papers.nips.cc/paper/1713-policy-gradient-methods-for-reinforcement-learning-with-function-approximation
54. Team, A.: Introducing claude 4. Tech. rep., Technical report, Anthropic/Stanford CRFM (2025)
55. Team, G.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. CoRR **abs/2507.06261** (2025). https://doi.org/10.48550/ARXIV.2507.06261, https://doi.org/10.48550/arXiv.2507.06261
56. 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. CoRR **abs/2506.07044** (2025). https://doi.org/10.48550/ARXIV.2506.07044, https://doi.org/10.48550/arXiv.2506.07044
57. Voigt, H., Sugamiya, Y., Lawonn, K., Zarrieß, S., Takanishi, A.: Llm-powered virtual patient agents for interactive clinical skills training with automated feedback. CoRR **abs/2508.13943** (2025). https://doi.org/10.48550/ARXIV.2508.13943, https://doi.org/10.48550/arXiv.2508.13943
58. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W.X., Wei, Z., Wen, J.: A survey on large language model based autonomous agents. CoRR **abs/2308.11432** (2023). https://doi.org/10.48550/ARXIV.2308.11432, https://doi.org/10.48550/arXiv.2308.11432
59. Wang, Y., Yang, Q., Zeng, Z., Ren, L., Liu, L., Peng, B., Cheng, H., He, X., Wang, K., Gao, J., Chen, W., Wang, S., Du, S.S., Shen, Y.: Reinforcement learning for reasoning in large language models with one training example. CoRR **abs/2504.20571** (2025). https://doi.org/10.48550/ARXIV.2504.20571, https://doi.org/10.48550/arXiv.2504.20571
60. Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., Chen, W.: Mmlu-pro: A more robust and challenging multi-task language understanding

benchmark. CoRR **abs/2406.01574** (2024). https://doi.org/10.48550/ARXIV.2406.01574, https://doi.org/10.48550/arXiv.2406.01574

61. Wang, Z., Wang, K., Wang, Q., Zhang, P., Li, L., Yang, Z., Jin, X., Yu, K., Nguyen, M.N., Liu, L., Gottlieb, E., Lu, Y., Cho, K., Wu, J., Fei-Fei, L., Wang, L., Choi, Y., Li, M.: RAGEN: understanding self-evolution in LLM agents via multi-turn reinforcement learning. CoRR **abs/2504.20073** (2025). https://doi.org/10.48550/ARXIV.2504.20073, https://doi.org/10.48550/arXiv.2504.20073
62. Yang, S., Zhao, H., Zhu, S., Zhou, G., Xu, H., Jia, Y., Zan, H.: Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. CoRR **abs/2308.03549** (2023). https://doi.org/10.48550/ARXIV.2308.03549, https://doi.org/10.48550/arXiv.2308.03549
63. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. CoRR **abs/2311.16502** (2023). https://doi.org/10.48550/ARXIV.2311.16502, https://doi.org/10.48550/arXiv.2311.16502
64. Yun, J., Sohn, J., Park, J., Kim, H., Tang, X., Shao, Y., Koo, Y., Ko, M., Chen, Q., Gerstein, M., Moor, M., Kang, J.: Med-prm: Medical reasoning models with stepwise, guideline-verified process rewards. CoRR **abs/2506.11474** (2025). https://doi.org/10.48550/ARXIV.2506.11474, https://doi.org/10.48550/arXiv.2506.11474
65. Zhan, R., Li, Y., Wang, Z., Qu, X., Liu, D., Shao, J., Wong, D.F., Cheng, Y.: Exgrpo: Learning to reason from experience. CoRR **abs/2510.02245** (2025). https://doi.org/10.48550/ARXIV.2510.02245, https://doi.org/10.48550/arXiv.2510.02245
66. Zhang, H., Chen, J., Jiang, F., Yu, F., Chen, Z., Chen, G., Li, J., Wu, X., Zhang, Z., Xiao, Q., Wan, X., Wang, B., Li, H.: Huatuogpt, towards taming language model to be a doctor. In: Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023. pp. 10859–10885. Association for Computational Linguistics (2023). https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.725, https://doi.org/10.18653/v1/2023.findings-emnlp.725
67. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: PMC-VQA: visual instruction tuning for medical visual question answering. CoRR **abs/2305.10415** (2023). https://doi.org/10.48550/ARXIV.2305.10415, https://doi.org/10.48550/arXiv.2305.10415
68. Zheng, H., Zhou, Y., Yan, T., Chen, D., Lu, H., Liao, W., He, T., Peng, P., Shen, J.: Clinical cognition alignment for gastrointestinal diagnosis with multimodal llms. CoRR **abs/2603.20698** (2026). https://doi.org/10.48550/ARXIV.2603.20698, https://doi.org/10.48550/arXiv.2603.20698
69. Zhou, C., Gong, Q., Zhu, J., Luan, H.: Research and application of large language models in healthcarecurrent development of large language models in the healthcare fielda framework for applying large language models and the opportunities and challenges of large language models in healthcare: A framework for applying large language models and the opportunities and challenges of large language models in healthcare. In: Proceedings of the 2023 4th International Symposium on Artificial Intelligence for Medicine Science, ISAIMS 2023, Chengdu, China, October 20-22, 2023. pp. 664–670. ACM (2023). https://doi.org/10.1145/3644116.3644226, https://doi.org/10.1145/3644116.3644226

70. Zhou, Y., Sheng, J., Wang, Q., Shen, J.: Compatibility-aware dynamic fine-tuning for large language models. arXiv preprint arXiv:2606.11206 (2026)
71. Zhou, Y., Song, L., Shen, J.: Improving medical large vision-language models with abnormal-aware feedback. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025. pp. 12994–13011. Association for Computational Linguistics (2025). https://doi.org/10.18653/V1/2025.ACL-LONG.636, https://doi.org/10.18653/v1/2025.acl-long.636
72. Zhou, Y., Song, L., Shen, J.: MAM: modular multi-agent framework for multimodal medical diagnosis via role-specialized collaboration. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025. Findings of ACL, vol. ACL 2025, pp. 25319–25333. Association for Computational Linguistics (2025). https://doi.org/10.18653/V1/2025.FINDINGS-ACL.1298, https://doi.org/10.18653/v1/2025.findings-acl.1298
73. Zhou, Y., Tao, W., Guo, Y., Shen, J.: World models meet language models: On the complementarity of concrete and abstract reasoning. arXiv preprint arXiv:2606.03603 (2026)
74. Zhou, Y., Zheng, H., Chen, D., Yang, H., Han, W., Shen, J.: From medical llms to versatile medical agents: A comprehensive survey. Authorea **2026**(0324) (2026). https://doi.org/10.22541/au.177437366.61663600/v1, https://www.authorea.com/doi/abs/10.22541/au.177437366.61663600/v1
75. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., Gao, Z., Cui, E., Wang, X., Cao, Y., Liu, Y., Wei, X., Zhang, H., Wang, H., Xu, W., Li, H., Wang, J., Deng, N., Li, S., He, Y., Jiang, T., Luo, J., Wang, Y., He, C., Shi, B., Zhang, X., Shao, W., He, J., Xiong, Y., Qu, W., Sun, P., Jiao, P., Lv, H., Wu, L., Zhang, K., Deng, H., Ge, J., Chen, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. CoRR **abs/2504.10479** (2025). https://doi.org/10.48550/ARXIV.2504.10479, https://doi.org/10.48550/arXiv.2504.10479
76. Zuo, Y., Qu, S., Li, Y., Chen, Z., Zhu, X., Hua, E., Zhang, K., Ding, N., Zhou, B.: Medxpertqa: Benchmarking expert-level medical reasoning and understanding. CoRR **abs/2501.18362** (2025). https://doi.org/10.48550/ARXIV.2501.18362, https://doi.org/10.48550/arXiv.2501.18362