

Clinical-R1: Empowering Large Language Models for Faithful and Comprehensive Reasoning with Clinical Objective Relative Policy Optimization

Boyang Gu^{1*}, Hongjian Zhou^{2*}, Bradley Max Segal², Jinge Wu³, Zeyu Cao⁴, Hantao Zhong⁴, Lei Clifton², Fenglin Liu², David A. Clifton^{2†}

¹Imperial College London

²University of Oxford

³University College London

⁴University of Cambridge

boyang.gu19@imperial.ac.uk, {Hongjian.zhou, bradley.segal, fenglin.liu}@eng.ox.ac.uk, jinge.wu.20@ucl.ac.uk, {zeyu.cao, hantao.zhong}@cl.cam.ac.uk, lei.clifton@ndph.ox.ac.uk, davidc@robots.ox.ac.uk

Abstract

Recent advances in large language models (LLMs) have shown strong reasoning capabilities through large-scale pre-training and post-training reinforcement learning, demonstrated by DeepSeek-R1. However, current post-training methods, such as Grouped Relative Policy Optimization (GRPO), mainly reward correctness, which is not aligned with the multi-dimensional objectives required in high-stakes fields such as medicine, where reasoning must also be faithful and comprehensive. We introduce Clinical-Objective Relative Policy Optimization (CRPO), a scalable, multi-objective, verifiable reinforcement learning method designed to align LLM post-training with clinical reasoning principles. CRPO integrates rule-based and verifiable reward signals that jointly optimize accuracy, faithfulness, and comprehensiveness without relying on human annotation. To demonstrate its effectiveness, we train Clinical-R1-3B, a 3B-parameter model for clinical reasoning. The experiments on three benchmarks demonstrate that our CRPO substantially improves reasoning on truthfulness and completeness over standard GRPO while maintaining comfortable accuracy enhancements. This framework provides a scalable pathway to align LLM reasoning with clinical objectives, enabling safer and more collaborative AI systems for healthcare while also highlighting the potential of multi-objective, verifiable RL methods in post-training scaling of LLMs for medical domains¹.

1 Introduction

Large language models (LLMs) (Brown et al. 2020; OpenAI 2023; Hurst et al. 2024; Jaech et al. 2024; Chowdhery et al. 2022; Google 2023; Touvron et al. 2023a,b) have achieved remarkable performance across a wide range of tasks, demonstrating strong generalization and reasoning abilities. As model scale increases, researchers have

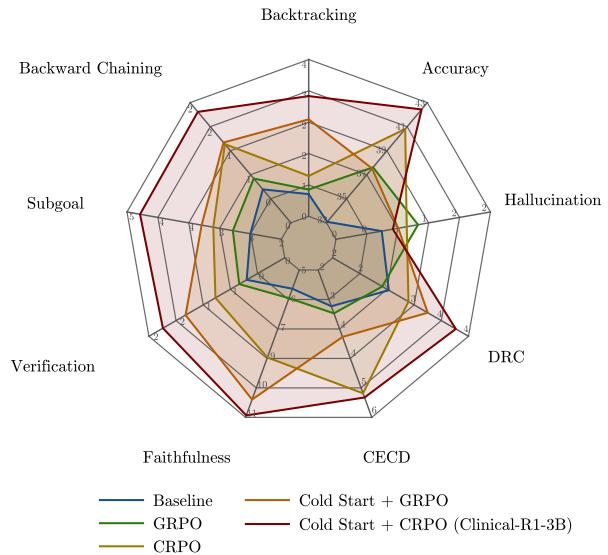


Figure 1: The accuracy, medical-faithfulness, and comprehensiveness evaluation across different methods. The result is averaged over 3 datasets (MedQA, MedMCQA, and MedXpertQA). For consistency of interpretation (higher is better), the Hallucination score is reported as $(100 - \text{Hallucination})$.

shifted focus from optimizing final responses to improving the reasoning process itself. Early methods introduced Chain-of-Thought (CoT) prompting and fine-tuning on reasoning data, while recent advancements in reinforcement learning (RL) have further enhanced reasoning capabilities during post-training (Rafailov et al. 2023; Ouyang et al. 2022; Shao et al. 2024; Guo et al. 2025).

Current popular RL-based algorithms, such as Proximal Policy Optimization (PPO) and Direct Preference Optimization (DPO) (Xu et al. 2024; Wang et al. 2024b; Rafailov et al. 2023), often face scalability and correctness challenges due to their reliance on human feedback (Rafailov et al. 2023;

*These authors contributed equally to this work.

†Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹Our data, models, and code are all publicly available on <https://github.com/BoyangGu1/Clinical-R1-3B>.

Ouyang et al. 2022). Reinforcement Learning from Verifiable Rewards (RLVR), and particularly Grouped Relative Policy Optimization (GRPO) (Shao et al. 2024), overcomes this limitation by replacing human preferences with explicit, rule-based rewards. GRPO has been shown to improve reasoning performance across mathematics and programming tasks, enabling models to discover complex reasoning behaviors through self-play without expert labeling. For instance, the DeepSeek-R1 model (Guo et al. 2025) employed RLVR to self-evolve complex reasoning patterns through training on logic puzzles, achieving versatile and advanced reasoning skills without relying on traditional techniques such as Monte Carlo Tree Search or Process Reward Models.

However, current GRPO formulations optimize for correctness alone, rewarding only the final answer rather than the reasoning process that leads to it. This single-objective design is misaligned with domains like medicine, where reasoning must also be faithful and comprehensive to achieve user trust and ensure clinical safety and regulatory compliance. Clinical reasoning requires the model not only to reach a correct conclusion but to provide verifiable, step-by-step justifications that clinicians can follow, evaluate and trust.

To address this limitation, we propose Clinical-objective Relative Policy Optimization (CRPO), a multi-objective extension of GRPO tailored for clinical reasoning. CRPO introduces rule-based and verifiable reward functions that jointly optimize for three objectives: accuracy, faithfulness and comprehensiveness. This approach enables models to develop reasoning processes that align with clinical expectations while maintaining scalability and training stability. We validate our method by training Clinical-R1-3B, a 3-billion-parameter model specialized in non-imaging clinical reasoning. Built upon a domain-distilled base model, Clinical-R1-3B is optimized using CRPO on multiple-choice clinical reasoning datasets. As shown in Figure 1, experimental results demonstrate that CRPO effectively encourages beneficial reasoning behaviors while suppressing irrelevant ones, achieving higher faithfulness and comprehensiveness compared to standard GRPO. Overall, our contributions are as follows:

- We design Clinical-objective Relative Policy Optimization (CRPO) specifically tailored for LLM post-training reinforcement learning in the medical domain, promoting faithfulness and reasoning comprehensiveness while improving accuracy, without needing human annotation.
- We introduce Clinical-R1-3B, a lightweight LLM optimized with CRPO for faithful and comprehensive clinical reasoning, bridging the gap between LLM reasoning and real-world clinical applications.
- We demonstrate through experiments on three benchmarks that Clinical-R1-3B improves the faithfulness and comprehensiveness of base models while achieving significant improvement in accuracy equivalent to popular methods such as GRPO, highlighting the potential of multi-objective, verifiable RL methods in training LLMs for high-stakes applications such as complex clinical decision support.

2 Related Work

2.1 Medical LLMs with Reasoning

There are many previous works focusing on Medical Large Language Models (LLMs) (Liu et al. 2025; Singhal et al. 2023; Yang et al. 2024b; Saab et al. 2024; Toma et al. 2023; Chen et al. 2023; Labrak et al. 2024) in enhancing reasoning and addressing limitations such as hallucinations and inaccuracy in complex tasks. One promising approach is to integrate medical knowledge from outside structured knowledge sources, such as knowledge graphs and retrieval systems. Wang et al. introduced JMLR, a model that jointly trains LLMs and retrieval systems to improve medical question-answering, achieving a notable reduction in hallucinations and training time compared to previous models (Wang et al. 2024a). Another possible method is to establish reasoning at test time. MedAdapter offers an efficient solution for adapting LLMs to biomedical applications without extensive computational resources via ranking multiple candidates at test time (Shi et al. 2024). Meanwhile, MedAgents facilitates zero-shot medical reasoning via a collaborative agent framework, generating reasoning iteratively (Tang et al. 2023). There are only a few works that focus on directly fine-tuning LLMs with medical knowledge in an RL manner. Chen et al. introduced HuatuoGPT-01, a medically focused model that utilizes Proximal Policy Optimization (PPO) for complex reasoning, highlighting the need for constant verification in healthcare settings (Chen et al. 2024). Pan et al. apply vanilla GRPO method on vision-language medical QA task and achieve substantial improvements (Pan et al. 2025). However, in almost all works, high-quality reasoning data or external knowledge is required for text-based medical QA tasks.

2.2 Reasoning with RL

Traditional RL-based reasoning methods are policy-based, such as PPO and Direct Preference Optimization (DPO) (Xu et al. 2024; Wang et al. 2024b; Rafailov et al. 2023). Cao et al. introduced DRLC to enhance RL by employing dense rewards generated through LLMs that can be used for PPO training later (Cao et al. 2024). DPO offers an alternative approach by directly optimizing language models to align with human preferences, where the model learns to distinguish between preferred and non-preferred responses (Rafailov et al. 2023). Besides policy-based RL, some works focus on value-based RL, especially Monte Carlo Tree Search (MCTS) (Qi et al. 2024; Guan et al. 2025). Qi et al. presented rStar, a mutual reasoning approach that enhances the problem-solving capabilities of small language models (SLMs) without fine-tuning, by employing a self-play mutual generation-discrimination process with MCTS at test time (Qi et al. 2024).

3 Methodology

The CRPO stage fine-tunes the model to produce final answers with reasoning under explicit, human-free reward signals. In detail, given a question q , CRPO samples a group of outputs $\{o_1, o_2, \dots, o_G\}$ from the updated policy π_θ and

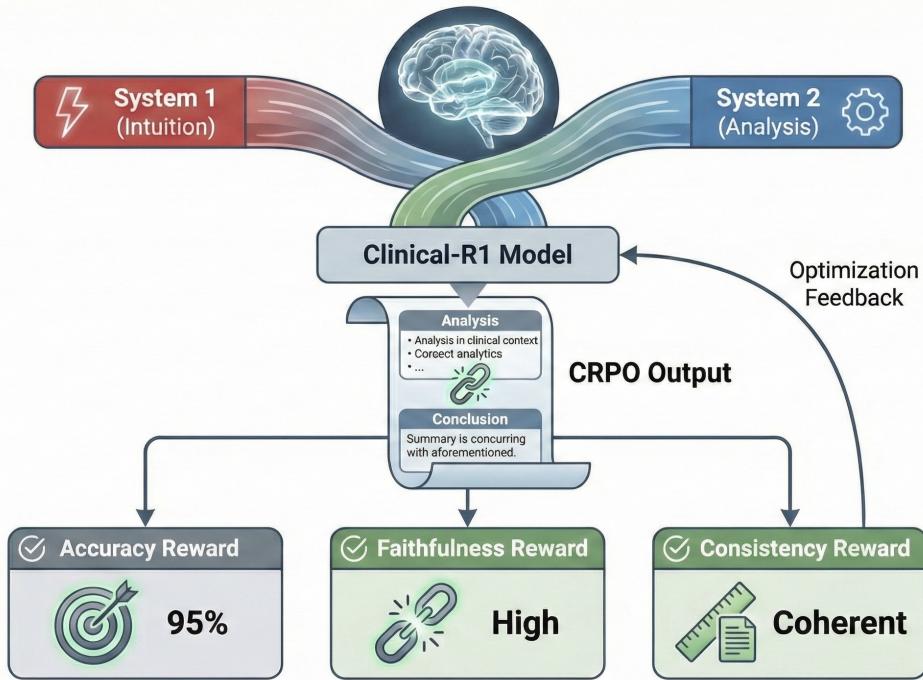


Figure 2: Overview of our Clinical-objective Relative Policy Optimization (CRPO) Design. The model is refined via on-policy CRPO with the above reward design.

then optimizes it by maximizing the following objective:

$$\mathcal{J}_{RPO}(\theta) = \mathbb{E}_{q \sim P(Q), \{\omega_i\}_{i=1}^G \sim \pi_\theta(O|q)} \left[\frac{1}{G} \sum_{i=1}^G (A_i - \beta \mathbb{D}_{KL}(\pi_\theta \| \pi_{ref})) \right], \quad (1)$$

where

$$\mathbb{D}_{KL}(\pi_\theta \| \pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1, \quad (2)$$

β is the KL-divergence regularization coefficient, π_{ref} is the reference model (in our case the original model untrained), and A_i is the advantage estimated, i.e.,

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (3)$$

CRPO allows our model to acquire transparent reasoning patterns while maintaining training stability and efficiency on limited computational resources. During inference, we aggregate multiple responses through majority voting to ensure robustness and filter out inconsistent reasoning paths.

Clinical Reasoning Reward Current post-training methods such as GRPO optimize only for correctness, without constraining how reasoning is formed. As a result, models often produce fragmented or shortcut explanations. This makes the reasoning process unreliable and hard to validate, especially in medicine, where faithfulness and completeness are as critical as accuracy. Clinicians, in contrast, rely on Dual Process Thinking (Djulbegovic et al. 2012):

an analytical process (System 2) that systematically evaluates hypotheses, and an intuitive process (System 1) that draws on experience and pattern recognition. Physicians often use System 1 to generate hypotheses and System 2 to verify or refine them, ensuring that intuitive insights remain grounded in analytical reasoning. To align LLMs with this cognitive structure, as shown in Figure 2, we introduce the Clinical Reasoning Reward, which enforces a structured reasoning format reflecting these two systems. The model must separate analytical reasoning and conclusion using $\langle dx \rangle$, $\langle /dx \rangle$, $\langle conclusion \rangle$, and $\langle /conclusion \rangle$ tags. The $\langle dx \rangle$ section captures hypothesis-driven reasoning, while the $\langle conclusion \rangle$ section synthesizes these insights into a conclusion. Additional reward is given when the conclusion explicitly references analytical elements from the dx section, promoting faithfulness and comprehensiveness.

Reward Design Figure 2 shows our designed rule-based reward that follows the convention created by GRPO with small variations. The reward is the summation of three parts: accuracy reward, clinical reasoning reward, and consistency reward.

- **Accuracy reward:** The accuracy reward calculates the correctness of the answer. In our case, the accuracy reward is determined by whether the model gives the correct choice in the multiple choices. If the model's answer is correct, the accuracy reward is 1. Otherwise, it is 0.
- **Clinical reasoning reward:** We require the model output to generate reasoning and the answer between the $\langle dx \rangle$, $\langle /dx \rangle$, $\langle conclusion \rangle$, and $\langle /conclusion \rangle$ tags. Within

the `<dx>`tags, we explicitly ask for a separation of an intuitive process (System 1) and an analytical process (System 2). If the response is in such format, a reward of 1 is given. To further encourage faithfulness and comprehensiveness, if the `<conclusion>`part cross-references the `<dx>`part, an additional reward of 0.5 is given. Otherwise, the clinical reasoning reward is 0.

- **Consistency reward:** GRPO-based methods may allow models to explore off-task responses or produce irrelevant tokens, such as content in other languages, especially when comprehensiveness is rewarded. To mitigate this, we define a consistency reward based on the ratio of effective tokens to total tokens in each response. This discourages off-rail exploration and maintains linguistic and contextual coherence throughout training.

Our final reward for one response o is calculated by

$$r(o) = k \cdot r_{\text{accuracy}}(o) + r_{\text{CR}}(o) + 0.5 \cdot r_{\text{consistency}}(o), \quad (4)$$

where k is a multiplier for estimating the importance of accuracy, so in total, the maximum reward a response can get is $k + 2$.

4 Experiments

4.1 Dataset

In this study, we utilize three medical QA datasets: MedQA (Jin et al. 2021), MedMCQA (Pal, Umapathi, and Sankarasubbu 2022), and MedXpertQA (Zuo et al. 2025). All three datasets are designed to evaluate models’ performance in multiple-choice question answering (MCQA) tasks in the medical domain.

MedQA (Jin et al. 2021) MedQA is derived from professional medical licensing examinations, including the United States Medical Licensing Examination (USMLE) and exams from mainland China and Taiwan. It comprises questions in three languages: English, simplified Chinese, and traditional Chinese. Each question is accompanied by four possible answer choices. The dataset covers a wide range of medical topics, requiring deep understanding and reasoning abilities. We will only use MedQA-English (12,723 questions) for our training and testing.

MedMCQA (Pal, Umapathi, and Sankarasubbu 2022) MedMCQA is a comprehensive dataset focused on multiple-choice questions from medical entrance examinations in India, specifically the All India Institute of Medical Sciences (AIIMS) and the National Eligibility cum Entrance Test Postgraduate (NEET PG). It includes over 194,000 questions. Each sample contains a question, correct answer(s), other options, and a detailed explanation, necessitating advanced language comprehension and reasoning skills.

MedXpertQA (Zuo et al. 2025) MedXpertQA is a challenging benchmark for expert-level medical reasoning. It comprises 4,460 questions spanning 17 specialties and 11 body systems, and includes text-only questions and multimodal questions with clinical images and structured patient data. The dataset undergoes rigorous filtering and augmentation, and is explicitly designed to test reasoning capacity

beyond factual recall. For our evaluation, we will use its text-only questions.

Together, these datasets provide complementary coverage of clinical knowledge, reasoning styles, and question difficulty. They serve as the foundation for training and evaluating our model’s ability to generate transparent, verifiable, and clinically grounded medical reasoning.

4.2 Baseline and Evaluation Metrics

We compare our Qwen2.5-3B-based Clinical-R1 with the original pretrained model with CoT prompting. The CoT setting uses the same reasoning template as the CRPO sampling prompt shown in Figure 2. This baseline establishes the performance of the base model before any distillation or reinforcement learning.

To evaluate the difference between CRPO and GRPO. We train two sets of models. The first set are trained with GRPO reward design, which only the final accuracy and the general thinking format are taken into consideration. The other set is trained with CRPO as discussed in Section 3. For each set, we have two models. One is the same as the baseline model, and the other is distilled by a stronger LLM, which we call the cold-start model.

There are some recent works discussing the influence of cognitive behaviors of the base model on the effectiveness of GRPO exploration (Gandhi et al. 2025). We extend this idea to the medical domain by evaluating the presence of cognitive reasoning patterns commonly associated with effective human problem solving as the degree of reasoning comprehensiveness. Specifically, we assess:

- **Backtracking**, where the model identifies an error in its reasoning and revisits earlier steps to revise its logic.
- **Answer Verification**, in which the model explicitly checks the consistency or plausibility of its final answer before concluding.
- **Subgoal Setting**, where intermediate objectives are introduced to structure the overall problem-solving process.
- **Backward-Chaining**, where reasoning is constructed in reverse from potential answers to the premises, mimicking diagnostic or hypothesis-driven inference.

In addition to cognitive behavior analysis, we further evaluate the medical-faithfulness of the generated reasoning from four medical-fact-centric perspectives. These include:

- **Faithfulness to Medical Knowledge**, which measures how many distinct, clinically relevant factual claims in the model’s reasoning are case-aligned, consistent with current standards of care, and factually accurate for the specific scenario.
- **Case-grounded Evidence Citation Density (CECD)**, which assesses the extent to which the reasoning explicitly links concrete patient-specific findings (e.g., vital signs, laboratory values, physical findings, imaging, or exposures) to medical inferences relevant to this particular case.

- **Distractor Rejection Coverage (DRC)**, which quantifies thoroughly the reasoning how explicitly and correctly rejects each incorrect answer choice with clinically valid justification.
- **Hallucination**, which detects unsupported or fabricated claims inconsistent with the question context or medical standards.

All of the above evaluations are done in an LLM-as-judge manner. Each of the above evaluations is performed by Llama-3.1-8B-Instruct (Grattafiori et al. 2024) and GPT5 using specifically designed prompts that count the number of distinct instances relevant to each dimension. We evaluate the cognitive behavior, content quality, and the accuracy of models’ responses with MedQA, MedMCQA, and MedXpertQA datasets.

4.3 Prompt and Training Parameters

In our experiment, we use MedQA for Cold Start and GRPO training. The test set of MedQA is treated as in-domain evaluation, while MedMCQA and MedXpertQA are treated as out-of-domain evaluation. The base model we use is Qwen2.5-3B-Instruct (Yang et al. 2024a). We utilize the Volcano Engine Reinforcement Learning (verl) framework for our Cold Start (SFT) and GRPO training (Sheng et al. 2024). All experiments are executed on 8xA6000 48GB machines.

Cold Start As discussed in Section 3, we distill DeepSeek-R1 on half of the MedQA training set questions (about 5,000) for 20 epochs as the cold-start dataset. With an early stopping evaluated on the validation set of MedQA, the fine-tuned model is trained for 13 epochs.

GRPO and CRPO For GRPO and CRPO training, our prompt is designed to guide the model to follow our desired output format for good initialization of training. For the cold-start base model, we train on the remaining half (about 5,000) of the MedQA training set after distillation for 20 epochs. The rollout number (G) is 5 as a compromise for the training time limit. For CRPO, the importance of accuracy coefficient $k = 10$.

Our final proposed model Clinical-R1-3B consists of two main stages: a cold-start initialization via supervised distillation and an on-policy CRPO optimization with rule-based verifiable rewards. The distillation stage provides a domain-aligned starting point by transferring reasoning traces from the strong teacher model DeepSeek-R1. Our analysis in Section 5.2 shows that the cold-start initialization successfully enhances the model’s ability of backtracking, answer verification, subgoal setting, and backward-chaining compared to the original model.

5 Results and Discussion

5.1 Results

As shown in Figure 3 and 4, across three benchmarks (MedQA in-domain; MedMCQA and MedXpertQA out-of-domain), CRPO-trained models match or surpass GRPO in answer accuracy. The cold-start + CRPO variant (Clinical-R1-3B) yielded the strongest competitive results. Majority

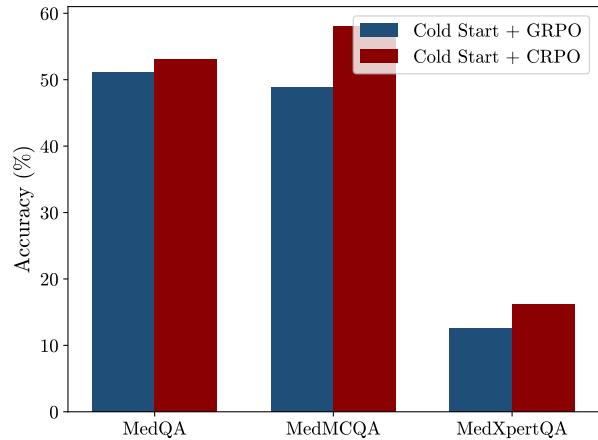


Figure 3: Accuracy Comparison with GRPO and CRPO.

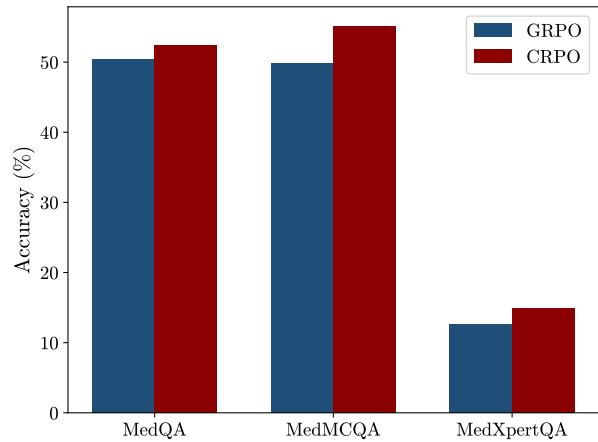


Figure 4: Accuracy Comparison with GRPO and CRPO.

voting adds little in already low-variance settings, suggesting that CRPO promotes self-consistent generations where ensembling offers limited marginal benefit (Table 1). The results show the effectiveness of our approach.

5.2 Cognitive Comprehensiveness

As shown in Table 1, CRPO consistently promotes clearer and more reliable clinical reasoning than GRPO, and a cold start initialization offers a stable starting point. The gains are most evident across the core cognitive behaviors below, where CRPO’s verifiable formatting and cross-referencing requirements ($\langle \text{dx} \rangle$ and $\langle \text{conclusion} \rangle$) encourage the model to separate presentation from etiology, surface intermediate structure, and perform explicit self-checks.

- **Backtracking.** With CRPO, the model is more likely to recognize a mistaken line of thought and revise it before finalizing an answer. In a lymphoma risk-factor case, GRPO prose tends to deviate from the causal exposure toward a correlated disease label, whereas CRPO’s structure prompts the model to notice that it has conflated *prior disease* with *causal exposure*, return to the earlier step, and correct the reasoning chain.

| Methods | Dataset: MedMCQA | | | | | | | | |
|--|------------------|-------------|-------------|--------------|--------------|-------------|-------------|----------------|--------------|
| | Backtracking | BC | Subgoal | Verification | Faithfulness | CECD | DRC | Hallucination↓ | Accuracy (%) |
| Baseline | 0.36 | 0.19 | 2.04 | 0.18 | 4.66 | 1.42 | 1.84 | 0.40 | 46.78 |
| w/ GRPO | 0.40 | 0.25 | 1.88 | 0.17 | 4.71 | 1.51 | 1.75 | 0.85 | 49.87 |
| w/ CRPO | 0.73 | 0.62 | 2.51 | 0.53 | 7.13 | 5.36 | 2.79 | 0.64 | 55.13 |
| w/ Cold Start + GRPO | 1.88 | 1.00 | 3.39 | 0.89 | 7.66 | 2.31 | 2.33 | 0.81 | 48.86 |
| w/ Cold Start + CRPO (Clinical-R1-3B) | 2.49 | 1.06 | 4.23 | 1.28 | 9.15 | 5.47 | 3.24 | 0.66 | 58.10 |

| Methods | Dataset: MedQA | | | | | | | | |
|--|----------------|-------------|-------------|--------------|--------------|-------------|-------------|----------------|--------------|
| | Backtracking | BC | Subgoal | Verification | Faithfulness | CECD | DRC | Hallucination↓ | Accuracy (%) |
| Baseline | 0.66 | 0.75 | 2.93 | 0.78 | 6.38 | 3.92 | 2.50 | 0.69 | 41.95 |
| w/ GRPO | 0.76 | 0.98 | 3.48 | 0.92 | 6.85 | 4.14 | 2.43 | 1.21 | 50.35 |
| w/ CRPO | 1.18 | 1.63 | 3.69 | 1.30 | 9.28 | 5.35 | 2.77 | 1.06 | 52.41 |
| w/ Cold Start + GRPO | 2.34 | 1.87 | 3.06 | 1.72 | 11.09 | 4.72 | 3.23 | 0.81 | 51.07 |
| w/ Cold Start + CRPO (Clinical-R1-3B) | 3.29 | 2.33 | 4.95 | 2.02 | 12.95 | 5.76 | 3.36 | 0.76 | 53.07 |

| Methods | Dataset: MedXpertQA | | | | | | | | |
|--|---------------------|-------------|-------------|--------------|--------------|-------------|-------------|----------------|--------------|
| | Backtracking | BC | Subgoal | Verification | Faithfulness | CECD | DRC | Hallucination↓ | Accuracy (%) |
| Baseline | 0.68 | 0.73 | 2.84 | 0.72 | 6.25 | 3.61 | 4.14 | 0.69 | 10.51 |
| w/ GRPO | 0.89 | 0.98 | 3.47 | 0.92 | 6.88 | 3.88 | 4.01 | 1.13 | 12.64 |
| w/ CRPO | 1.18 | 1.68 | 3.77 | 1.23 | 9.27 | 5.33 | 3.79 | 1.05 | 14.88 |
| w/ Cold Start + GRPO | 3.20 | 1.14 | 4.20 | 1.77 | 12.07 | 4.42 | 4.64 | 0.80 | 12.64 |
| w/ Cold Start + CRPO (Clinical-R1-3B) | 3.42 | 2.15 | 5.06 | 2.08 | 10.66 | 5.15 | 4.84 | 0.79 | 16.14 |

Table 1: Medical reasoning and accuracy results across prompts and optimization methods. Bold numbers indicate best-performing methods per dataset (lower is better for Hallucination). BC, CECD, and DRC represent Backward Chaining, Case-grounded Evidence Citation Density, and Distractor Rejection Coverage, respectively. As we can see, our proposed Clinical-R1-3B achieves the best performance on most cases.

- **Backward-chaining.** CRPO promotes reasoning that begins with a candidate conclusion and then works backward to evidence that must appear in $\langle \text{dx} \rangle$. In the same lymphoma scenario, CRPO treats “radiation exposure” as a hypothesis and then requires supporting facts tied to the case (treatment history, pathologic subtype), rather than making loose associations between travel or sex and risk. This top-down constraint reduces leaps from conclusion to justification.
- **Subgoal setting.** Under GRPO’s correctness-only optimization, explanations are often compressed, omitting necessary intermediate steps. CRPO’s formatting requirements preserve a concise yet stepwise plan: shortlist plausible risk categories, separate presentation from risk, map each option to a risk type, then synthesize. The lymphoma example illustrates this decomposition: presentation is recognized as a finding, prior malignancy as background, and treatment exposure as the etiologic driver.
- **Verification.** Because CRPO rewards conclusions that explicitly cite elements from $\langle \text{dx} \rangle$, the model habitually performs a final consistency check: Does the conclusion rest on case-grounded evidence rather than on context or coincidence? In the lymphoma case, this verification pass

filters out non-causal distractors (travel, sex, location of nodes) and reaffirms treatment exposure as the risk factor.

CRPO is the key enhancement of clinically reliable reasoning: its verifiable, multi-objective rewards convert model capacity into separating findings, background, and causal exposures, and requiring that conclusions cite the evidence on which they rest. This structure suppresses two GRPO failure modes: treating a presenting feature as a risk factor and elevating a prior diagnosis over the true causal exposure. A cold start provides stronger initialization (improved terminology and patterns), but without CRPO’s constraints it does not consistently prevent these errors. With CRPO, reasoning remains faithful to the case, systematic in rejecting distractors, and anchored to explicit, verifiable evidence.

5.3 Medical Faithfulness

We evaluate the factual and clinical quality of generated reasoning using four dimensions: faithfulness, case-grounded evidence citation density (CECD), distractor rejection coverage (DRC), and hallucination. Together these metrics reflect whether a model reasons accurately, cites case-relevant

evidence, rejects incorrect hypotheses, and minimizes unsupported statements.

As shown in Table 1, CRPO consistently yields the strongest medical faithfulness across datasets. The improvements are most pronounced in CECD and DRC, where the CRPO-trained system produces explanations that repeatedly point back to concrete tokens and provide concise case-specific refutations. Clinical-R1-3B (cold start + CRPO) attains the best overall profile, indicating that verifiable, multi-objective rewards are essential for aligning reasoning with patient-specific evidence and medical standards.

The qualitative contrast is clear. Under GRPO, answers often rely on broad associations without citing the quoted findings, overlooking anchors such as “every 45 days” and “160 mg/dL,” or drifting into taxonomic imprecision like calling ovarian hyperthecosis a subtype to be used interchangeably. In another example, GRPO focuses on a background label (“previous breast cancer”) instead of the causal exposure explicitly present in the history (“previous radiation therapy”), despite pathologic cues like “centroblastic and immunoblastic cell presence.” In contrast, CRPO formats $\langle dx \rangle$ as a ledger of case facts and requires $\langle conclusion \rangle$ to cite them. Distractors are then rejected with brief, evidence-bound rationales.

Hallucination behavior follows the same pattern. Cold start, by itself, encourages longer but less organized answer and can overgeneralize, occasionally introducing unsupported claims or loosened taxonomy. CRPO suppresses this tendency by making unsupported additions costly: conclusions must point to sentences already established in $\langle dx \rangle$, which reduces language drift and discourages invented labs, staging, or etiologies. GRPO alone reduces some errors tied to final-answer pressure but still permits generic dismissals and ungrounded phrasing when explicit evidence links are not rewarded.

In summary, cold start strengthens the initializer, but CRPO regulates medical faithfulness. Verifiable, multi-objective rewards convert raw capacity into an auditable workflow that consistently ties claims to quoted case facts, broadens case-grounded citations, expands targeted distractor rejection, and reduces unsupported content. Their combination in Clinical-R1-3B produces reasoning that is faithful to medical knowledge and reliably anchored to the patient at hand.

5.4 Response Length Analysis

We observe a characteristic length pattern across training stages. Cold start produces long, fully enumerated chains. GRPO then aggressively compresses responses. CRPO settles in between, yielding concise yet sufficiently structured explanations. This reflects CRPO’s ability to prune redundancy while preserving clinically necessary steps.

The trend is visible both in the stabilization of token counts during CRPO training (Figure 5). Superficial segments are progressively removed while essential reasoning is retained, indicating a shift from verbose exploration to compact reasoning.

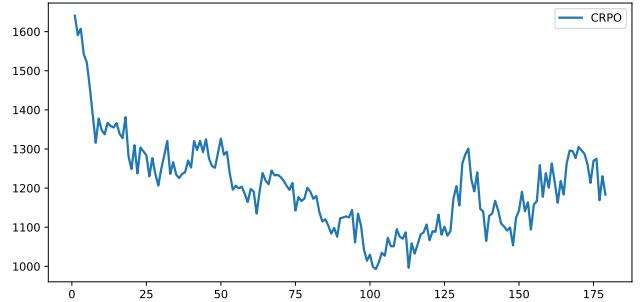


Figure 5: Response length vs training step (MedQA).

5.5 Case Study

Figure 6 shows a case example for cold-start + GRPO and cold-start + CRPO. For this particular case, in GRPO, the model frequently conflates disease labels with causal exposures, jumping from “prior breast cancer” to secondary malignancy risk without separating the disease label from the treatment exposure. Distractors such as sex, anatomic site, or travel are rarely preemptively negated, and mis-selection of “previous breast cancer” is common.

In Cold-start + GRPO, the model more often outlines alternatives and attempts eliminations, reflecting inherited structure. However, the final mapping remains vulnerable to high-salience yet non-causal labels; radiation exposure is not consistently isolated from the disease label. Errors decrease relative to pure GRPO but persist under strong distractors.

In CRPO, a three-way partition is established before scoring options: causal exposure (radiation) versus disease label/comorbidity (breast cancer) versus presentation/epidemiologic context (axillary involvement, sex, travel). The exposure timeline and directionality are aligned to the outcome, and non-causal features are proactively rejected. The model selects previous radiation therapy and provides explicit negative evidence for alternatives.

In Cold-start + CRPO, subgoals are articulated (categorize options, align case facts, eliminate distractors), and brief verification confirms that exposure precedes outcome. The resulting chain is compact and auditable: radiation is the causal exposure; breast cancer is a label; sex/site/travel constitute presentation or weak-context features. This case illustrates the general conclusion that cold start improves structural coverage, while CRPO, with or without cold start, provides stronger cognitive comprehensiveness and medical faithfulness.

6 Conclusion and Limitation

In this work, we presented Clinical-objective Relative Policy Optimization (CRPO), a scalable, multi-objective, and verifiable RL framework that aligns post-training with clinical reasoning principles. By combining rule-based rewards for accuracy, faithfulness, and comprehensiveness to enforce a lightweight reasoning answer ($\langle dx \rangle$, $\langle conclusion \rangle$). CRPO improves the quality of the reasoning process without relying on human annotation. Trained with CRPO, our 3B-parameter Clinical-R1-3B achieves stronger medical faith-

Question:

A 77-year-old woman presents with fever, malaise, and night sweats. She recently noted enlargement of axillary lymph nodes that she examines weekly... She has a remote history of breast cancer in her 60s treated with radiation and chemotherapy; extensive travel to Africa; and a 30 pack-year smoking history. On exam, several axillary nodes are palpable with a large, non-tender right axillary mass (10×8 cm). FNAB shows "centroblastic and immunoblastic cell presence, suspicious for NHL–DLBCL variant."

Which risk factor is responsible?

- A) Travel to Africa B) Axillary lymph node involvement C) Previous radiation therapy D) Female gender E) Previous breast cancer

Ground truth: C**Cold Start + GRPO Answer**

<think>“...77-year-old woman with fever, malaise, night sweats... enlarged axillary lymph nodes... FNA suggests DLBCL.”

“Option A: Travel to Africa... but radiation therapy is a known risk factor...”

“Option B: part of the presentation, not a risk.”

“Option C: Previous radiation therapy... a big risk factor.”

“Option E: Previous breast cancer... known risk factor for lymphoma...”

(weighs options, then conflates prior disease with causation) (...) </think>

<answer>Discusses prior breast cancer treated with radiation, but concludes **E (previous breast cancer)** as the “risk factor” rather than the treatment exposure.

E </answer>

Cold Start + CRPO Answer

<dx>(System 1: Intuitive) “...Age 77 with B-symptoms (fever, night sweats, malaise) and a bulky 10×8 cm axillary mass; FNAB “centroblastic/immunoblastic” → DLBCL plausible.” Notes history of breast cancer *treated with radiation/chemotherapy*, Africa travel, and smoking... Key risk candidates arise: *prior radiation*, immunosuppression/HIV/EBV, certain infections; gender is unlikely...

(System 2: Analytical) Risk vs. finding clarification: (B) *axillary involvement* = presentation, not a risk.

(A) *Travel to Africa*: can relate to endemic EBV/Burkitt... but *not the pertinent causal risk* for this cytology.

(D) *Female gender*: not established as a DLBCL risk.

(E) *Previous breast cancer*: the disease history itself is *not causal*; the **treatment exposure** is.

(C) **Previous radiation therapy**: consistent with radiation-induced lymphomagenesis (DNA damage, secondary NHL), matches this patient’s prior breast-cancer *radiation*...

Synthesizes: distinguish presentation vs. etiology; map each option to risk type; elevate **C** as the causal exposure while demoting non-causal context (A/E) and mere findings (B)... (...) </dx>

<conclusion>Most relevant risk factor for this DLBCL presentation is the patient’s **previous radiation therapy** (treatment exposure), not prior malignancy itself.

C </conclusion>

Figure 6: Case study on a DLBCL risk-factor question. Ellipses (...) indicate trimmed content.

fulness and reasoning comprehensiveness than GRPO while maintaining higher answer accuracy across three benchmarks MedQA, MedMCQA, and MedXpertQA (Table 1). These gains are accompanied by desirable cognitive behaviors (backtracking, backward-chaining, subgoal setting, and verification), suggesting that verifiable, multi-objective optimization is an effective pathway for safer, more collaborative clinical LLMs.

Despite these promising results, several limitations remain. First, CRPO optimization can be unstable, particularly without a cold-start initialization, due to rapid KL-divergence growth between the policy and reference models. This instability may lead to inefficient training or partial convergence. Second, cognitive comprehensive and medical faithfulness evaluation rely on automatic annotation by another LLM, which may not fully align with human judgment. As a result, human evaluation is required for stronger

validation. Third, our experiments are conducted on base models not pretrained on medical corpora, limiting the representational depth of domain knowledge.

Future work should explore more stable variants of CRPO (e.g., adaptive or off-policy updates), introduce human-in-the-loop evaluations, and extend the method to stronger medical backbones. We hope this line of research contributes toward building smaller yet more interpretable and trustworthy medical reasoning systems.

Acknowledgements

DAC was funded by an NIHR Research Professorship; a Royal Academy of Engineering Research Chair; and the InnoHK Hong Kong Centre for Cerebro-cardiovascular Engineering (COCHE); and was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and the Pandemic Sciences Institute

at the University of Oxford. The Applied Digital Health (ADH) group at the Nuffield Department of Primary Care Health Sciences is supported by the National Institute for Health and Care Research (NIHR) Applied Research Collaboration Oxford and Thames Valley at Oxford Health NHS Foundation Trust. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. FL was funded by the Clarendon Fund and the Magdalen Graduate Scholarship. HZ was funded by the Clarendon Fund, the Department of Engineering Science Studentship, and the Frederick Brodckhues Scholarship. BMS is funded by the Rhodes Trust under the Rhodes scholarship. We sincerely thank all the editors and reviewers for their constructive comments and suggestions that substantially improved this paper.

References

- Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *Annual Conference on Neural Information Processing Systems*.
- Cao, M.; Shu, L.; Yu, L.; Zhu, Y.; Wickers, N.; Liu, Y.; and Meng, L. 2024. Drlc: Reinforcement learning with dense rewards from llm critic. *arXiv e-prints*, arXiv:2401.
- Chen, J.; Cai, Z.; Ji, K.; Wang, X.; Liu, W.; Wang, R.; Hou, J.; and Wang, B. 2024. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*.
- Chen, Z.; Cano, A. H.; Romanou, A.; Bonnet, A.; Matoba, K.; Salvi, F.; Pagliardini, M.; Fan, S.; Köpf, A.; Mottashami, A.; et al. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. *arXiv preprint arXiv:2311.16079*.
- Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Djulbegovic, B.; Hozo, I.; Beckstead, J.; Tsalatsanis, A.; and Pauker, S. G. 2012. Dual processing model of medical decision-making. *BMC medical informatics and Decision Making*, 12(1): 94.
- Gandhi, K.; Chakravarthy, A.; Singh, A.; Lile, N.; and Goodman, N. D. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*.
- Google. 2023. Bard: A Generative Artificial Intelligence Chatbot. <https://gemini.google.com>.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Guan, X.; Zhang, L. L.; Liu, Y.; Shang, N.; Sun, Y.; Zhu, Y.; Yang, F.; and Yang, M. 2025. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking. *arXiv preprint arXiv:2501.04519*.
- Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.
- Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14): 6421.
- Labrak, Y.; Bazoge, A.; Morin, E.; Gourraud, P.-A.; Rouvier, M.; and Dufour, R. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.
- Liu, F.; Zhou, H.; Gu, B.; Zou, X.; Huang, J.; Wu, J.; Li, Y.; Chen, S. S.; Hua, Y.; Zhou, P.; et al. 2025. Application of large language models in medicine. *Nature Reviews Biotechnology*, 1–20.
- OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.
- Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, 248–260. PMLR.
- Pan, J.; Liu, C.; Wu, J.; Liu, F.; Zhu, J.; Li, H. B.; Chen, C.; Ouyang, C.; and Rueckert, D. 2025. Medvilm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 337–347. Springer.
- Qi, Z.; Ma, M.; Xu, J.; Zhang, L. L.; Yang, F.; and Yang, M. 2024. Mutual reasoning makes smaller llms stronger problem-solvers. *arXiv preprint arXiv:2408.06195*.
- Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.
- Saab, K.; Tu, T.; Weng, W.-H.; Tanno, R.; Stutz, D.; Wulczyn, E.; Zhang, F.; Strother, T.; Park, C.; Vedadi, E.; et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.
- Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

- Sheng, G.; Zhang, C.; Ye, Z.; Wu, X.; Zhang, W.; Zhang, R.; Peng, Y.; Lin, H.; and Wu, C. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*.
- Shi, W.; Xu, R.; Zhuang, Y.; Yu, Y.; Sun, H.; Wu, H.; Yang, C.; and Wang, M. D. 2024. Medadapter: Efficient test-time adaptation of large language models towards medical reasoning. *arXiv preprint arXiv:2405.03000*.
- Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; Schaekermann, M.; Wang, A.; Amin, M.; Lachgar, S.; Mansfield, P.; Prakash, S.; Green, B.; Dominowska, E.; y Arcas, B. A.; Tomasev, N.; Liu, Y.; Wong, R.; Semturs, C.; Mahdavi, S. S.; Barral, J.; Webster, D.; Corrado, G. S.; Matias, Y.; Azizi, S.; Karthikesalingam, A.; and Natarajan, V. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. *arXiv preprint arXiv:2305.09617*.
- Tang, X.; Zou, A.; Zhang, Z.; Li, Z.; Zhao, Y.; Zhang, X.; Cohan, A.; and Gerstein, M. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.
- Toma, A.; Lawler, P. R.; Ba, J.; Krishnan, R. G.; Rubin, B. B.; and Wang, B. 2023. Clinical Camel: An Open-Source Expert-Level Medical Language Model with Dialogue-Based Knowledge Encoding. *arXiv preprint arXiv:2305.12031*.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Wang, J.; Yang, Z.; Yao, Z.; and Yu, H. 2024a. Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. *arXiv preprint arXiv:2402.17887*.
- Wang, Z.; Bi, B.; Pentyala, S. K.; Ramnath, K.; Chaudhuri, S.; Mehrotra, S.; Mao, X.-B.; Asur, S.; et al. 2024b. A comprehensive survey of LLM alignment techniques: RLHF, RLAIF, PPO, DPO and more. *arXiv preprint arXiv:2407.16216*.
- Xu, S.; Fu, W.; Gao, J.; Ye, W.; Liu, W.; Mei, Z.; Wang, G.; Yu, C.; and Wu, Y. 2024. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*.
- Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Yang, L.; Xu, S.; Sellergren, A.; Kohlberger, T.; Zhou, Y.; Ktena, I.; Kiraly, A.; Ahmed, F.; Hormozdiari, F.; Jaroensri, T.; et al. 2024b. Advancing Multimodal Medical Capabilities of Gemini. *arXiv preprint arXiv:2405.03162*.
- Zuo, Y.; Qu, S.; Li, Y.; Chen, Z.; Zhu, X.; Hua, E.; Zhang, K.; Ding, N.; and Zhou, B. 2025. MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. *arXiv preprint arXiv:2501.18362*.