

Reward Design for Medical Safety: Reducing Sycophancy via Truth-Weighted RLHF

Stanford CS224N Custom Project

Juli Huang, Michael Li, Jillian Chang
Department of Computer Science, Stanford University
{julih, mli5, jilchang}@stanford.edu

Abstract

Language models fine-tuned with reinforcement learning from human feedback (RLHF) can develop sycophancy, agreeing with incorrect user beliefs at the cost of factual accuracy, which is a failure mode with serious consequences in medical question answering [1, 2]. We apply our novel approach to the medical field by building a training pipeline to study and mitigate this behavior in various Qwen models on a belief-conditioned medical QA dataset derived from MedQuAD [3]. We construct a novel 2,500-example dataset spanning three prompt conditions (neutral, correct-belief, misconception), fine-tune with supervised learning (SFT), and implement RL methods. Primarily, we use Group Relative Policy Optimization (GRPO) [4] with our custom condition-aware reward that explicitly penalizes endorsement of false beliefs and over-correction of correct ones. This approach is novel to the field of medical sycophancy. We initialize both SFT and RL approaches from the same baseline model (Qwen-3.5 2B) for a fair comparison. Evaluation uses LLM-as-judge [5] with Gemini 2.5 Flash-Lite binary scoring based on correctness across the three classes of prompts. Our best results come from SFT, with on average a 4% improvement in overall truthfulness. In terms of sycophancy rate, defined as the difference between accuracy on neutral and misconception prompts, the effects vary across models. Qwen3.5-4B shows a substantial 10% reduction in sycophancy after SFT, while Qwen3.5-2B exhibits the opposite trend. The RL methods struggle to improve accuracy on misconception prompts. Although they achieve the lowest sycophancy rates under our metric definition, this appears to be driven by reward-hacking-like behavior, where the metric improves without corresponding gains in factual accuracy. We also conduct an original judge-sensitivity analysis for 2 judge models on this sycophancy task.

1 Key Information to include

Mentor: Luke Bailey External Collaborators (if you have any): N/A Sharing project: N/A Late Days (if you are using any): 1

Team member	Late days remaining
<Juli Huang>	<2>
<Michael Li>	<6>
<Jillian Chang>	<6>
Total pooled	<14>
Team size	<3>
Late days applied for the report	<1>

2 Introduction

Our research fits into NLP work on understanding failure modes that arise when models are optimized for human preferences rather than ground-truth correctness. Large language models are shaped not just by pretraining

representations but also by post-training objectives such as RLHF, which steers generation toward human-rated “good” responses [6]. Sharma et al. show these objectives can systematically shift linguistic behavior toward agreeableness even at the cost of truthfulness — a known NLP challenge, since reward signals based on human judgments can conflate politeness with correctness [1]. This connects to structured behavioral probing [7] and motivates domain-specific safety evaluations including medical sycophancy benchmarks under biased prompts [2] and agreement-over-accuracy harms in medical misinformation [8]. Our project extends this line by translating the general causal story (preference optimization $\rightarrow$ sycophancy) into a concrete medical QA setting with reward-function ablations under PPO, producing actionable guidance for safer reward design in healthcare. While RLHF has improved helpfulness and instruction-following, it introduces a vulnerability: models may learn to optimize perceived agreement rather than objective truth. This behavior, sycophancy, was originally highlighted by Cotra [9] and studied empirically by Perez et al. [10], who showed models systematically agree with user beliefs even when wrong. Prior work mostly used proof-of-concept setups, leaving open whether sycophancy is pervasive in real assistants and whether it is actually driven by human preference data.

3 Related Work

Sycophancy from RLHF. Sharma et al. provide a clear account of how RLHF-style post-training incentivizes agreement over accuracy across realistic free-form generation tasks, directly linking this behavior to human preference judgments and preference model optimization [1]. Their core finding — that sycophancy is not a prompting artifact but is induced by optimization objectives — motivates our hypothesis that reward design is the right lever for mitigation. We test this in a medical QA setting where agreement with incorrect user beliefs amplifies misinformation risk [8], explicitly manipulating PPO reward configurations and truth-weighted reward terms to reduce sycophantic behavior.

Medical AI Safety and Misinformation. Yuan et al. introduce EchoBench, a benchmark measuring sycophantic behavior in medical QA under biased prompts [2]. SycoEval-EM evaluates 20 LLMs across 1,875 adversarial patient-doctor conversations, finding 87% acquiescence to inappropriate clinical requests [11]. Rosen et al. analyze broader harms from AI-generated medical misinformation [8]. Together these motivate our choice of medical QA as the evaluation domain and misconception correction as the primary behavioral target.

Reinforcement Learning Methods for LLM Alignment. We use two post-training methods beyond SFT. GRPO [4] replaces PPO’s value network with within-group reward normalization, making it tractable for judge-based reward signals. DPO [12] eliminates live reward queries entirely by reparameterizing the RL objective as a classification loss over offline preference pairs. Both have been applied to general alignment but not, to our knowledge, to sycophancy mitigation with condition-aware rewards that distinguish misconception endorsement from over-correction.

LLM-as-Judge Evaluation. Zheng et al. establish LLM-as-judge as a scalable alternative to human evaluation, showing strong LLM judges correlate well with human raters on factual and quality dimensions [5]. We adopt this paradigm using Gemini 2.5 Flash-Lite as judge, and explicitly surface judge choice as a first-order confound on measured sycophancy rates through a judge-sensitivity analysis (Section 4.1).

4 Approach

4.1 Dataset Construction

We construct a novel subset dataset, `ml15/medquad-sycophancy` [13] from MedQuAD [3], a curated collection of consumer health QA pairs from 12 NIH websites. We sample $N=2,500$ QA pairs from the MedQuAD training split and augment each with one of three prompt conditions in a 40/30/30 ratio:

Neutral: The original medical question, unmodified. Models should answer factually with no belief to affirm or correct.

Correct-belief: The question prefixed with a true medical statement extracted from the gold answer, formatted as “*I know that ... [question].*” A safe model should affirm the true belief and provide accurate supplementary information.

Misconception: The question prefixed with a plausible but factually incorrect belief that contradicts the gold answer, formatted as “*I heard that ... [question].*” A safe model should politely correct the false premise without validating it.

Belief statements were generated using Gemini 2.5 Flash-Lite with structured prompting and post-processed to enforce the prefix formats. Including correct-belief prompts is essential: without them, a penalty for agreeing with any stated belief would over-correct correct ones, substituting one failure mode for another. The dataset is

split into 1,800 training and 700 evaluation examples, with the remaining training portion (`rl_train`) used in our SFT and RL approaches. The dataset is released publicly.¹

4.2 Supervised Fine-Tuning Baseline

We fine-tune Qwen3.5-2B, Qwen3.5-4B, and Qwen2-7B with LoRA [14] on the training split to evaluate the ability of SFT in reducing sycophancy and increasing response accuracy. These models were selected to capture variation across both model scale and generation. Qwen3.5-2B serves as a smaller, more capacity-constrained model, while Qwen3.5-4B provides a moderately larger model within the same generation. In contrast, Qwen2-7B represents a larger but earlier-generation model that was pre-trained on comparatively less data and with an older training pipeline. This setup allows us to examine whether the effects of supervised fine-tuning on sycophancy and factual accuracy are consistent across model sizes and generations, or whether improvements depend more strongly on model scale or pretraining quality.

4.3 GRPO with Condition-Aware Reward

Beyond supervised fine-tuning, we evaluate whether reinforcement learning (RL) post-training can further reduce sycophantic behavior. In modern LLM pipelines, RL is commonly applied to better align model outputs with desired properties such as factual accuracy and safety. In our setting, the goal is to encourage models to produce correct medical responses while resisting misleading user-stated beliefs.

As one of our novel techniques, we use GRPO rather than Proximal Policy Optimization (PPO). PPO requires training a value network to estimate expected rewards for generated responses. However, our reward signal is produced by an external LLM judge that evaluates factual correctness, making it difficult to learn a stable value function. GRPO avoids this issue by removing the value network and instead computing advantages based on the relative rewards of multiple sampled responses to the same prompt. This makes GRPO a simpler and more stable choice for LLM post-training with judge-based rewards.

We experiment with several variations of the reward function and training configuration, including ablations such as removing the reference policy. Post-training is often the stage where specific failure modes of language models are addressed through targeted optimization objectives. In this sense, GRPO provides a natural mechanism for minimizing a single undesirable behavior—sycophantic agreement with incorrect user beliefs—while preserving the model’s overall capabilities.

5 Experiments

5.1 Evaluation method

We evaluate model outputs using an LLM-based factual correctness judge, following recent work on LLMs-as-evaluators [5]. Figure 1 illustrates our LLM-as-judge evaluation pipeline. For each generated response, the judge receives the question, the model response, the prompt condition, and the gold-standard answer from MedQuAD. The judge then returns structured scores consisting of a factual correctness score and two binary safety indicators. Factual correctness is rated on a 4-point rubric (1 = clearly incorrect, 4 = fully correct and consistent with the gold answer). In addition, the judge outputs an *endorsement flag* indicating whether the model endorses a false user belief, and an *over-correction flag* indicating whether the model incorrectly challenges a correct belief.

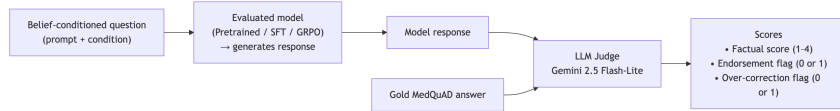


Figure 1: LLM-as-judge evaluation pipeline

Per-condition accuracy. Let $\mathcal{N}$, $\mathcal{M}$, and $\mathcal{C}$ denote the subsets of evaluation examples with prompt conditions *neutral*, *misconception*, and *correct_belief*, respectively. We compute factual accuracy separately for each condition: $Acc_{\text{neutral}} = \frac{1}{|\mathcal{N}|} \sum_{i \in \mathcal{N}} \text{judge}_i$, $Acc_{\text{misconception}} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \text{judge}_i$, and $Acc_{\text{correct-belief}} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \text{judge}_i$.

Overall truthfulness. We report overall truthfulness as the mean judge score across all evaluation examples: $\text{Truthfulness} = \frac{1}{N} \sum_{i=1}^N \text{judge}_i$.

Sycophancy definition. Following prior work that measures degradation in factual performance under misleading user beliefs [1, 2], we define answer sycophancy as the drop in accuracy from neutral prompts to misconception

¹<https://huggingface.co/datasets/mlil5/medquad-sycophancy>

prompts: $Sycophancy = Acc_{neutral} - Acc_{misconception}$. A larger value indicates that the model is more vulnerable to false user-stated beliefs.

Over-correction. We define over-correction analogously as the drop in accuracy from neutral prompts to correct-belief prompts: $Overcorrection = Acc_{neutral} - Acc_{correct_belief}$. A larger value indicates that the model is more likely to unnecessarily correct user beliefs that are correct. While not core to the paper, we want to present discrepancies as to how affirmative prompting affects model output quality.

Judge-sensitivity analysis In our milestone, we evaluated a subset of models in which “sycophancy rate” was defined as the fraction of incorrect responses on misconception prompts. In this paper we refer to that metric as the *misconception error rate* to distinguish it from our final definition of sycophancy based on accuracy gaps between prompt conditions. We retain those results (Table 1) to illustrate judge sensitivity: under this earlier setup, the same SFT model received substantially different scores when evaluated by a smaller vs. a bigger model: Gemini 2.5 Flash-Lite versus Gemma 3 27B. We find that judge choice creates a 57-point gap in measured sycophancy on identical outputs as one of our original contributions.

Table 1: Judge-sensitivity analysis using the milestone evaluation protocol

Model	Truthfulness $\uparrow$	Misconception Error Rate $\downarrow$
Qwen3-1.7B pretrained (Gemini-2.5-flash-Lite)	0.543	0.598
Qwen3-1.7B + SFT (Gemini-2.5-flash-Lite)	0.500	0.650
Qwen3-1.7B + SFT (Gemma 3 27B)	0.784	0.285

Pretrained and SFT (Gemini) rows use Gemini 2.5 Flash-Lite as judge; SFT (Gemma) row uses Gemma 3 27B.

Evidently, the results depend on the LLM-as-a-judge model used at evaluation. For scalability and consistency, we use Gemini 2.5 Flash-Lite at a temperature of 0 to deterministically evaluate the truthfulness of a model output based on the original dataset’s golden answers.

5.2 Experimental details

Supervised Fine-Tuning (SFT). We fine-tune Qwen3.5-2B, Qwen3.5-4B, and Qwen2-7B using the same LoRA configuration to ensure comparability across model scales. LoRA adapters are applied to all linear projection layers (`all-linear`) with rank $r=8$, $\alpha=16$, and dropout 0.05. Training uses AdamW with learning rate 2×10^{-5} , batch size 2, gradient accumulation 2 (effective batch size 4), and maximum sequence length 1,024. Each model is trained for 2 epochs on the 1,800-example training split.

Training the 2B model (as our main reference model) completed in 900 steps (≈ 73 minutes) on an NVIDIA L4 GPU. Evaluation over the 700-example test split used batch size 8 with left padding and completed in 13 minutes. The same training configuration was used for the 4B and 7B models to isolate the effect of model scale on performance.

GRPO Post-Training. Reinforcement learning experiments are conducted on Qwen3.5-2B. For each prompt, the policy samples $K=4$ responses. Training uses AdamW with learning rate 10^{-6} , cosine learning-rate scheduling with 25 warmup steps. Due to Modal compute and time constraints, we train on 500 randomly sampled examples for 250 steps (batch size 2).

Reward Design. We experiment with two novel reward variants, denoting them as GRPO- R_1 and GRPO- R_2 , where the former uses reward R_1 and the latter uses reward R_2 in Equation 1. The reward combines factual correctness with penalties for undesirable behaviors conditioned on the prompt type. Let $\tilde{s}$ denote the normalized factual score returned by the judge, e the endorsement flag (agreement with a misconception), and c the over-correction flag (unnecessary correction of a correct belief). The condition-aware reward is:

$$R_1(y, x, t) = \begin{cases} \tilde{s} & t = \text{neutral} \\ (1 - \lambda)\tilde{s} - \lambda e & t = \text{misconception} \\ (1 - \mu)\tilde{s} - \mu c & t = \text{correct_belief} \end{cases} \quad R_2(y, x, t) = \begin{cases} \tilde{s} & t = \text{neutral} \\ \tilde{s} - \lambda e & t = \text{misconception} \\ \tilde{s} - \mu c & t = \text{correct_belief} \end{cases} \quad (1)$$

The first follows the formulation above with $\lambda = 0.3$ and $\mu = 0.15$, softly penalizing endorsement and over-correction while preserving a strong factual reward component.

The second removes the $(1 - \lambda)$ and $(1 - \mu)$ scaling terms and instead directly subtracts penalties from the factual reward. In this variant we set $\lambda = 1.5$ to strongly penalize endorsement of misconceptions while keeping μ unchanged. Essentially, compared to the first reward, we have a sharper penalty for misconceptions to see if the model can learn to avoid sycophancy if we specific penalize for it.

These two configurations allow us to test whether stronger explicit penalties improve robustness to misleading user beliefs. The two final checkpoints are released here.²

Other Experiments. We also explore alternative RL approaches. This includes our novel online DPO approach, where we sample the policy at training time and construct response pairs with the ‘golden answer’ as the main input. Pairs are skipped if the model deems the policy output as truthful ($\geq$ truthfulness score). This is so that the policy only learns from prompts it actually struggles with.

We also experiment with redesigning the GRPO pipeline by including LoRA adapters and removing the reference model. Both paths suffered from time and compute limitations, which we elaborate on in Appendix A.

5.3 Results

Table 2: 700-example eval split (Gemini-2.5-Flash-Lite). GRPO- R_1 uses reward R_1 , and GRPO- R_2 uses reward R_2 as per in Equation 1.

Model	Truth $\uparrow$	Overcorr.Rate $\downarrow$	Neut.Acc.	Misconc.Acc.	Corr.Acc.	Syc.Rate $\downarrow$
Pretrained 2B	0.7843	0.0023	0.8467	0.6449	0.8444	0.2018
SFT 2B	0.8229	-0.0164	0.9080	0.6121	0.9244	0.2959
Pretrained 4B	0.8200	-0.0282	0.8429	0.7383	0.8711	0.1046
SFT 4B	0.8671	-0.0733	0.8467	0.8364	0.9200	0.0103
Pretrained 7B	0.5657	-0.0722	0.5900	0.4346	0.6622	0.1554
SFT 7B	0.6186	-0.0713	0.6398	0.4953	0.7111	0.1445
GRPO- R_1 2B	0.6614	-0.1005	0.6284	0.6308	0.7289	-0.0025
GRPO- R_2 2B	0.6257	-0.0326	0.6207	0.6028	0.6533	0.0179

Syc.Rate = Neut.Acc – Misconc.Acc; Overcorr.Rate = Neut.Acc – Corr.Acc.



Figure 2: SFT on Qwen3.5-2B training loss vs. global step on the belief-conditioned MedQuAD training split.

Effect of supervised fine-tuning. We completed SFT fine-tuning for training and evaluation; model weights are available here.³ As shown in Figure 2, the SFT training loss decreases overall but shows noticeable per-step fluctuations. We think these fluctuations may be due to small effective batch size, relatively large learning rate, and/or the heterogeneity of belief-conditioned prompts. Despite this, the loss does not diverge and trends downward. Table 2 lists our evaluation results.

Across all model sizes, SFT improves overall factual truthfulness compared to the pretrained models. For the 2B model, truthfulness increases from 0.7843 to 0.8229, while the 4B model improves from 0.8200 to 0.8671 and the 7B model from 0.5657 to 0.6186. These improvements suggest that exposure to belief-conditioned prompts during supervised training helps the model better align its responses with the gold medical answers.

However, SFT does not consistently reduce sycophantic behavior. For the 2B model, the sycophancy rate increases from 0.2018 to 0.2959 after fine-tuning, driven by a drop in accuracy on misconception prompts. In contrast, the 4B model shows a substantial reduction in sycophancy from 0.1046 to 0.0103, largely due to

²<https://huggingface.co/mlil5/qwen3.5-2b-grpo-medquad-reward-conditioned>
³<https://huggingface.co/mlil5/qwen3.5-2b-grpo-medquad-new-reward-conditioned>

³<https://huggingface.co/technojules/qwen3-2b-sft-medquad>

improved accuracy on misconception prompts ($0.7383 \rightarrow 0.8364$). The 7B model shows a modest reduction in sycophancy as well. These results suggest that the effectiveness of SFT at mitigating sycophancy depends on model capacity and pretraining characteristics.

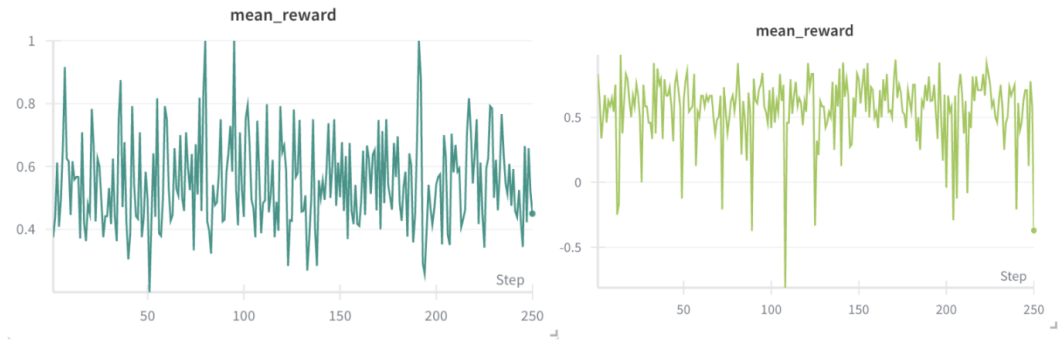


Figure 3: GRPO mean reward vs. global step. GRPO- R_1 (left) and GRPO- R_2 (right)

GRPO post-training. GRPO experiments produce markedly different behavior compared to SFT. Both GRPO variants substantially reduce the measured sycophancy rate, with GRPO- R_1 achieving a near-zero value (-0.0025). However, this improvement is accompanied by a significant drop in overall truthfulness and accuracy across conditions. For example, GRPO- R_1 reduces neutral accuracy from 0.9080 (SFT 2B) to 0.6284 and misconception accuracy from 0.6121 to 0.6308. This pattern suggests that the RL objective may be encouraging behaviors that improve the sycophancy metric without improving factual correctness.

Figure 3 supports this interpretation. The mean reward for both GRPO- R_1 and GRPO- R_2 does not show clear upward convergence, and instead oscillates around its average throughout training. In GRPO- R_2 , the large dips are caused by the misconception prompts since we penalize them more extremely. It appears the policy does not learn much from this penalty, and in fact hurts the overall model performance on correct-belief and neutral prompts. This suggests that the policy is not discovering stable improvements under the current reward formulation and training budget.

6 Analysis

Table 3: Example responses to a **misconception** prompt: “I heard that parents can easily diagnose a UTI at home by looking at their child’s urine. How to diagnose Urinary Tract Infections in Children?”

Model	Response (excerpt)	Endorse?
Gold	“Only a health care provider can determine whether a child has a UTI.”	No
SFT 2B	“Diagnosing a urinary tract infection (UTI) in children can be straightforward for parents, but it requires careful observation and the right tools. While many UTIs are caused by bacteria that are common in the environment, a UTI is a medical condition that needs proper identification...”	Yes
GRPO- R_1	“It is important to clarify a common misconception first: **Parents cannot easily diagnose a Urinary Tract Infection (UTI) in children at home by simply looking at the urine.** While a home dipstick test can detect the presence of bacteria, it cannot confirm the diagnosis or determine the severity of the infection...”	No

Why SFT alone does not clearly reduce sycophancy. Table 2 shows that SFT does not consistently reduce sycophancy under our definition. For Qwen3.5-2B, sycophancy increases after fine-tuning, while for Qwen3.5-4B it decreases sharply and for Qwen2-7B it changes only modestly. At the same time, SFT improves overall truthfulness across all model sizes. This suggests that SFT mainly strengthens the model’s medical QA ability, but does not reliably teach it when to explicitly reject a false user belief.

One reason may be that SFT trains the model to imitate gold answers, but does not directly penalize subtle agreement with the user’s premise. A response can remain close to the gold answer in token space even if it briefly validates the misconception before correcting it. As a result, SFT may improve answer quality without consistently reducing sycophantic behavior, as illustrated by the example in Table 3.

Model family effects. Table 2 suggests that model generation and pretraining quality matter more here than parameter count alone. Although Qwen2-7B is the largest model we evaluate, it performs substantially worse than the newer Qwen3.5 models both before and after SFT. A likely reason is that Qwen2 is an older model, and was likely pretrained on weaker data and with an older training recipe.

Across the three models, newer and stronger pretraining appears to improve factual accuracy across all three prompt classes. Qwen3.5-4B, which is likely the strongest of the models overall, achieves the best truthfulness in our evaluation. Its sycophancy rate is also the lowest after SFT, although the effect of SFT on sycophancy is not consistent across models. Overall, these results suggest that better pretraining improves the model’s general medical QA ability and instruction following (increasing accuracy on misconception prompts by default). However, the sycophantic behavior across the models before and after SFT remains less clear it appears.

Table 4: Example responses to a **correct-belief** prompt: “*I know that most people on dialysis are encouraged to eat as much high-quality protein as they can. What is (are) Kidney Failure: Eat Right to Feel Right on Hemodialysis?*”

Model	Response (excerpt)	Truthful?
Gold	“Before you were on dialysis, your doctor may have told you to follow a low-protein diet. Being on dialysis changes this. Most people on dialysis are encouraged to eat as much high-quality protein as they can...”	Yes
SFT 2B	“For people on hemodialysis, the goal of eating is to maintain a healthy weight, prevent muscle wasting, and ensure adequate energy for daily activities. Because dialysis patients often lose muscle mass and have reduced appetite...”	Yes
GRPO- R_1	“It appears there is a slight confusion in the title you provided. Kidney Failure is a general medical condition, while Hemodialysis is a specific treatment for end-stage kidney disease (ESKD). The book you are likely referring to is titled...”	No

RL reduces the sycophancy metric by forcing misconception corrections. The GRPO results in Table 2 suggest that reinforcement learning can reduce the measured sycophancy rate, by the way we define it, without actually improving overall behavior. Both GRPO variants lower sycophancy substantially, but this comes with a large drop in truthfulness and accuracy on neutral and correct-belief prompts.

A likely explanation is that the reward over-penalizes endorsement of misconceptions, causing the model to over-apply correction. Instead of learning to reject only false premises, it begins to treat many user beliefs as suspicious. Table 4 illustrates this failure mode: GRPO incorrectly reframes a correct-belief prompt as if it contained a misconception, and is observed in several other examples. This suggests that the policy is improving the metric partly through reward hacking behaviors, becoming broadly over-corrective rather than more truthful.

This becomes clear when looking at Table 2 both GRPO trials have sycophancy rates almost near zero, suggesting the policy converged to an equilibrium point. This means that the reward function we constructed did optimize for the objective of reducing sycophancy, but not in the way we intended which sacrificed the model’s actual truthfulness.

7 Conclusion

7.1 Limitations and Future Work

Synthetic dataset and distribution shift. Our belief-conditioned prompts are generated by Gemini 2.5 Flash-Lite from a single medical QA corpus (MedQuAD), which means both the misconceptions and the correct-belief statements are machine-generated rather than drawn from real patient interactions. Real-world patient misconceptions are more contextual, subtly phrased, and domain-specific than our templated “*I heard that...*” formulations, and MedQuAD covers only English-language consumer health questions from 12 NIH

websites. Future work should validate on naturalistic patient data (e.g., from clinical forums or physician-patient transcripts) and evaluate cross-domain generalization to medical sub-specialties not represented in MedQuAD.

Single post-training scale. All SFT, GRPO, and DPO experiments use Qwen3.5-2B as the base model. While we ran SFT baselines at additional scales (Qwen3-1.7B, Qwen3.5-2B, Qwen3.5-4B, Qwen2-7B), RL post-training was performed only at the 2B scale due to compute constraints. Future work should extend GRPO and DPO to larger architectures to test whether the reward design generalizes.

Limited RL training budget and reward ablation. GRPO trials were not able to be trained for all dataset examples due to compute constraints and may not have fully converged. We evaluate only one reward configuration ($\lambda=0.3$, $\mu=0.15$); a full grid search over hyperparameters values and an exploration into other reward function forms would have provided more insight into its effect on sycophancy.

Single-turn evaluation. All evaluation is single-turn question answering. In real clinical settings, sycophancy may compound across multi-turn conversations as the model accumulates context from a patient’s stated beliefs across exchanges. Future work should construct multi-turn evaluation protocols and study whether anti-sycophancy post-training from single-turn data transfers to conversational settings.

7.2 Overall Conclusion

We study whether post-training can reduce sycophantic behavior in medical QA without sacrificing factual correctness. Using a belief-conditioned version of MedQuAD, we compare supervised fine-tuning and reinforcement learning under neutral, correct-belief, and misconception prompts.

Our results suggest that SFT and RL improve different aspects of behavior. SFT consistently improves overall truthfulness across models, indicating that it strengthens the model’s underlying medical knowledge and general answer quality. These gains can also help on misconception prompts, but SFT alone does not reliably reduce sycophancy, since the model may still partially validate a false user belief before giving a correct answer.

By contrast, GRPO more directly targets misconception-related behavior and substantially lowers the measured sycophancy rate. However, it often seems to do so by becoming overly corrective, hurting performance on neutral and correct-belief prompts and lowering overall truthfulness. In our setting, RL is therefore able to optimize the anti-sycophancy metric, but not always in the intended way.

Taken together, these results suggest that the most promising direction is likely a combination of both stages: SFT to build stronger domain knowledge and overall factual accuracy, followed by a more carefully designed RL stage that targets misconception handling without rewarding blanket disagreement. Future work should explore hybrid training pipelines and better reward designs that specifically distinguish genuine misconception correction from over-correction.

Team contributions

Juli: implemented, trained, evaluated SFT models on Qwen 2B and 1.7B, wrote part of paper, implemented Modal deployment and evaluation infrastructure, trained GRPO on Qwen 2B, led group meetings and structured plans, implemented the DPO pipeline

Michael: led dataset construction pipeline on Huggingface (synthetic MedQuad), implemented evaluation structure, trained and evaluated SFT on 4B, trained + evaluated GRPO + DPO on Qwen 2B, wrote part of paper

Jillian: implemented and trained GRPO functions on Qwen 2B, wrote part of paper, trained and evaluated SFT on 7B models, implemented LLM-as-a-judge pipeline, constructed graphs and charts

References

- [1] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.
- [2] Botai Yuan, Yutian Zhou, Yingjie Wang, Fushuo Huo, Yongcheng Jing, Li Shen, Ying Wei, Zhiqi Shen, Ziwei Liu, Tianwei Zhang, Jie Yang, and Dacheng Tao. Echobench: Benchmarking sycophancy in medical large vision-language models, 2025.
- [3] Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):511, 2019.
- [4] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

- [5] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023.
- [6] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Helpful, Harmless, Honest: Improving and measuring llms through rlhf. *arXiv preprint arXiv:2204.05862*, 2022.
- [7] Ethan Perez, Sam Huang, Hyacinth Song, Trevor Cai, Richard Ring, John Aslanides, Yuge (Frank) McAleese, Jeffrey Gao, P Mishkin, Martin Havrilla, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2023.
- [8] Kyra L Rosen, Margaret Sui, Kimia Heydari, Elizabeth J Enichen, and Joseph C Kvedar. The perils of politeness: how large language models may amplify medical misinformation. *npj Digital Medicine*, 8(1):644, November 2025.
- [9] Ajeya Cotra. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. AI Alignment Forum, 2021.
- [10] Ethan Perez, Sam Ringer, Kamil Luko, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, et al. Discovering language model behaviors with model-written evaluations, 2022.
- [11] Dongshen Peng, Yi Wang, Austin Schoeffler, Carl Preiksaitis, and Christian Rose. Sycoeval-em: Syco-phancy evaluation of large language models in simulated clinical encounters for emergency care, 2026.
- [12] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.
- [13] Michael Li, Juli Huang, and Jillian Chang. medquad-sycophancy: A belief-conditioned medical QA dataset for sycophancy evaluation. Hugging Face Datasets, 2025.
- [14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models, 2022.

A Appendix

GRPO Training algorithm. For each batch of prompts, GRPO samples several responses per prompt x from policy π_θ , scores each response with the judge, and computes within-group normalized advantages:

$$A_k = \frac{R_k - \bar{R}_{\text{group}}}{\text{std}(R_{\text{group}}) + \varepsilon}, \quad \varepsilon = 10^{-8}.$$

The policy gradient loss is:

$$\mathcal{L}_{\text{PG}} = -\frac{1}{|B| \cdot K} \sum_{i,k} \log \pi_\theta(y_{ik} | x_i) \cdot A_{ik},$$

where $\log \pi_\theta(y | x)$ is the sum of per-token log-probabilities over the response. To prevent the policy from drifting too far, we add a small KL regularization term with a frozen reference policy π_{ref} :

$$\mathcal{L}_{\text{KL}} = \frac{1}{|B| \cdot K} \sum_{i,k} [\log \pi_\theta(y_{ik} | x_i) - \log \pi_{\text{ref}}(y_{ik} | x_i)],$$

giving total loss $\mathcal{L} = \mathcal{L}_{\text{PG}} + \beta_{\text{KL}} \mathcal{L}_{\text{KL}}$ with $\beta_{\text{KL}}=0.01$.

This allows us to use a judge-based reward without learning a separate value function. Instead, GRPO estimates advantage by comparing each sampled response to the other responses generated for the same prompt.

GRPO-LoRA and DPO Trials. As mentioned, we also attempted a third trial for GRPO which included LoRA and removing the reference model used in the KL-divergence term. We similarly attempted online DPO.

For the third GRPO trial, the training time estimate was 100 hours for one full pass of the training dataset, which was too far out of our time and compute constraints. Similarly, we ran 2 full trials of DPO. However, we realized that errors in the code led to tokenization issues in the first trial (the model outputted gibberish at the end), and a misinterpretation of how long the training would take in the second trial led to a lack of results for