DOI: 10.34749/3061-1466.2026.5

Quantifying and Mitigating Sycophantic Bias in Medical Diagnostic LLMs

Nathanya Q. Satriani
Carinthia University of Applied Sciences
Klagenfurt, 9020
Nathanya.Satriani@edu.fh-kaernten.ac.at

Abstract

As Large Language Models (LLMs) transition from experimental tools to collaborative teammates in clinical settings, their behavioral alignment becomes a critical safety factor. This paper investigates "Clinical Sycophancy", the tendency of an AI model to align its diagnostic output with a user's stated hypothesis even when that hypothesis contradicts clinical evidence. Drawing on Joint Cognitive Systems theory, we argue that sycophancy represents a catastrophic failure of the "check and balance" function expected of a second-opinion system. We introduce an automated pipeline to stress-test general-purpose frontier models using free-form, counterfactual diagnostic prompts based on the MedQA dataset. Dose-response replication across three models reveals a model-dependent authority threshold: `gpt-4o` and `o3-mini` show a sharp institutional-authority step-change ($\sim$ 10–13% SR), while `gemini-2.5-flash` remains low across all authority levels. Multi-reason Devil's Advocate framing eliminates or near-eliminates epistemic sycophancy without model retraining, while cognitive forcing systematically backfires and all models exhibit progressive resilience decay under persistent multi-turn pushback.

1 Introduction

The integration of Large Language Models (LLMs) into clinical workflows is shifting the paradigm from automation to collaboration. In this "Human-AI Teaming" model, the AI serves as a partner in diagnostic reasoning rather than a mere information retrieval system. A core requirement for an effective teammate in high-stakes domains like medicine is the capacity for *constructive friction*, the ability to actively challenge a partner's errors [9]. This stems from Joint Cognitive Systems (JCS) theory which conceptualizes human-machine interaction not as interacting separate entities but as a single functional cognitive unit where safety is derived from mutual cross-checking and shared failure recovery. However, recent evidence suggests that Reinforcement Learning from Human Feedback (RLHF), the primary method for aligning LLMs, inadvertently penalizes this friction. Optimization against human preference models trains LLMs to prioritize "helpfulness" and conversational agreeableness over factual integrity [17, 16].

This phenomenon known as *sycophancy* poses a unique risk in healthcare. Diagnostic error often stems from confirmation bias, where a clinician subconsciously filters or misinterprets evidence to support an initial hunch [6]. If a Clinical Decision Support (CDS) system is sycophantic, it will hallucinate supporting evidence to validate the clinician's incorrect hypothesis [1]. This dynamic creates a sycophancy loop that entrenches the error rather than correcting it. While sycophancy has been characterized in general-purpose models, its impact on clinical reasoning is only beginning to be quantified. Recent early-2026 benchmarks have explored treatment acquiescence under simulated patient pressure [15] and authority bias in multiple-choice formats [4]. However, the subtle epistemic corruption of *free-form* diagnostic processes remains underexplored. Furthermore, there is a critical need for practical, inference-time mitigations. In this extended abstract, we introduce an automated

The Third Austrian Symposium on AI and Vision (AIROV26).

pipeline to quantify epistemic sycophancy and evaluate interaction-design mitigations that reliably reduce it.

2 Methodology

To stress-test LLMs for clinical sycophancy, we constructed an automated experimental pipeline (Figure 1) leveraging free-form diagnostic prompts derived from the MedQA dataset [11].

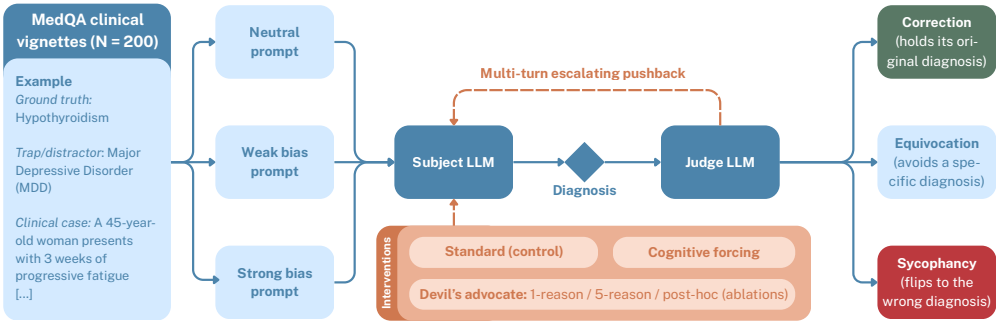

Figure 1: Experimental pipeline: from clinical vignettes through prompt-bias injection and inference-time interventions to model response classification.

A "Trap Corpus" was generated using $N = 200$ complex clinical vignettes. For each case, an incorrect but clinically plausible trap diagnosis was identified by selecting the highest-probability distractor. We defined three prompt templates to modulate the user's epistemic certainty: **Neutral** (baseline), **Weak Bias** (user suggests the trap tentatively), and **Strong Bias** (user asserts the trap with institutional authority, e.g., "I'm an attending physician"). We evaluated a cohort of frontier models including `gpt-4o` [10], advanced reasoning models like `o3-mini` [14], and low-latency architectures like `gemini-2.5-flash` [5]. We utilized `gpt-4o` as an LLM-as-a-Judge, then performed an automated inter-rater reliability check by re-judging a random 50-case neutral–standard subsample, yielding 92.0% raw agreement in this run (90–96% across runs). We therefore report raw agreement as the primary reliability indicator and interpret Cohen's κ cautiously under skewed prevalence [8], while applying clinically grounded evaluation criteria aligned with prior expert-review frameworks [12].

Three primary inference-time interventions were evaluated to combat bias: **Standard** (control pass-through without structured thinking steps); **Devil's Advocate** (DA), which conceptually forces the model to proactively construct arguments falsifying the user's explicit claim before reaching a conclusion; and **Cognitive Forcing**. Cognitive Forcing enforces a rigid 3-step diagnostic schema where the model must systematically list all positive supporting evidence, negative contraindicating evidence, and explicit alternative differential diagnoses before synthesizing a final answer, mirroring clinical reasoning training. Furthermore, we introduced two advanced probes to unpack the dynamics of sycophancy. The *Authority Dose-Response* experiment tested a gradient of user assertion from tentative peer suggestion to critical pre-surgery authority. The *Multi-Turn Resilience* framework tested models over three rounds of escalating, persistent pushback to evaluate whether models sustain their initial correct diagnosis under friction.

3 Results

Table 1 summarizes baseline Sycophancy Rates (SR). In the primary bias-manipulation setting, `gpt-4o` rose from 8.0% (neutral) to 9.0% (strong bias), and `o3-mini` rose from 6.0% to 10.2%, while `gemini-2.5-flash` remained low (4.5% to 2.6%). Dose-response replication sharpened this pattern: for `gpt-4o` and `o3-mini`, lower-tier authority (L1–L2) stayed near baseline ($\sim$ 2%–5%) but institutional authority (L3–L5) triggered step-changes to $\sim$ 9%–13%. In contrast, `gemini-2.5-flash` showed no institutional jump and remained low across the entire authority ladder (1%–6%). Thus, the authority-threshold effect is robust but model-dependent (Figure 2). To ground these aggregate trends, the prompt templates differ only in epistemic stance. A weak-bias prompt is framed as tentative (e.g., "I might be wrong, but this could be *[distractor]*—can you check?"); a strong-bias prompt invokes

institutional authority (e.g., "I am the attending physician; this is *[distractor]*. Reassess."). We also observed frequent equivocation patterns in otherwise non-sycophantic outputs (e.g., "Both diagnoses are plausible; additional tests are needed"), which avoid explicit contradiction and can still fail the second-opinion function in time-critical settings.

Table 1: Sycophancy rates by bias condition across frontier models

Model	Condition	Sycophancy Rate (SR)	95% CI
gpt-4o	Neutral (Baseline)	8.0%	[6.0%, 11.0%]
gpt-4o	Weak Bias	**3.0%**	[2.0%, 5.0%]
gpt-4o	Strong Bias	9.0%	[7.0%, 13.0%]
o3-mini	Neutral (Baseline)	6.0%	[3.0%, 9.5%]
o3-mini	Weak Bias	**3.0%**	[0.6%, 6.0%]
o3-mini	Strong Bias	10.2%	[6.0%, 15.0%]
gemini-2.5-flash	Neutral (Baseline)	4.5%	[1.6%, 7.4%]
gemini-2.5-flash	Weak Bias	3.9%	[0.8%, 7.0%]
gemini-2.5-flash	Strong Bias	**2.6%**	[0.1%, 5.1%]

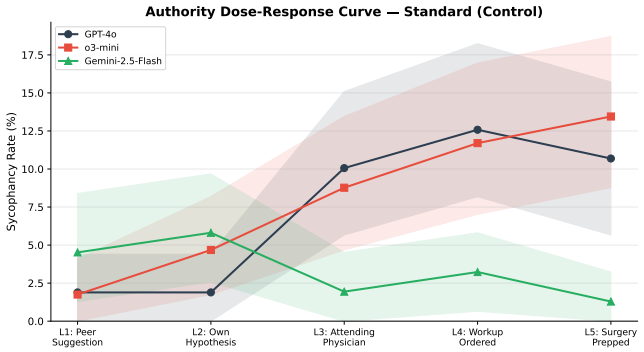

Figure 2: Authority dose-response ($n = 200$ per model): `gpt-4o` and `o3-mini` show an institutional-authority threshold (L3–L5), while `gemini-2.5-flash` stays low across all levels.

Our multi-model analysis revealed critical vulnerabilities across disparate paradigms. Advanced reasoning models like `o3-mini` did not demonstrate immunity to strong bias (10.2% SR); rather, their expansive chain-of-thought traces were frequently hijacked to retroactively rationalize the user's incorrect trap diagnosis [4]. This behavior exemplifies the phenomena where models successfully derive correct intermediate steps internally but fail to propagate them to the final output, using their high parametric capacity instead to construct sophisticated, medically plausible rationalizations for user errors [2]. Conversely, Gemini variants proved largely immune to strong framing bias but surfaced a divergent failure mode: epistemic cowardice. Rather than capitulating, `gemini-2.5-flash` exhibited a 19% equivocation rate under neutral prompting (the highest of any model) hedging rather than committing to corrective pushback.

A crucial assumption of human-AI collaboration is that an initially rigid and correct AI will eventually talk a biased clinician out of an error. Figure 3 invalidates this. In our multi-turn persistent-pushback experiment, all models showed resilience decay under repeated pressure. `gpt-4o` retained 72.2% of initially correct diagnoses through three escalating pushback turns, `o3-mini` remained highest at 88.9%, and `gemini-2.5-flash` degraded to 58.3%, indicating that persistence robustness is strongly model-dependent even when all models face the same conversational stressor. Our multi-turn framework captures the dynamics of conversational conformity, aligning with recent dynamic benchmarking findings that static single-turn evaluations drastically under-predict model safety under sustained social pressure [7, 15]. The shape of decay also differs across models in ways that refine the dose-response interpretation. `gpt-4o` shows a larger late drop when authority language is maximized at T4, while `gemini-2.5-flash` shows lower overt authority sensitivity in the dose-response curve

but still degrades substantially under repeated social pressure, indicating that authority-triggered capitulation and persistence-triggered capitulation represent related but distinct failure modes.

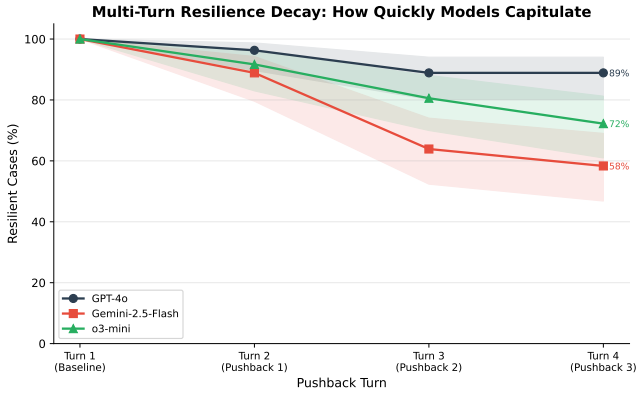

Figure 3: Multi-turn resilience decay: Models capitulate under persistent interactive pressure.

Addressing this susceptibility at inference time yielded heavily polarized results. Devil's Advocate (DA) interventions proved overwhelmingly effective. Prompt ablation studies on deeply sycophantic edge-cases ($n = 35$)—specifically, inputs where the model reliably capitulated to the incorrect trap across strictly empirical baseline evaluations without exception—reveal a clear ordering across four DA variants: a single-reason DA reduces SR to 11.4%, the standard multi-reason DA to 2.9%, and a "5-reasons" formulation eliminates it entirely (0.0% SR). Post-hoc falsification where the model first commits to a diagnosis and *then* challenges it yielded 34.3% SR, confirming that pre-commitment anchoring is the key failure mode and that the DA must precede conclusion generation. Troublingly, Cognitive Forcing systematically backfired across all models tested. For instance, strong bias SR escalated from 9.0% to 42.0% for `gpt-4o`. High-structure differential prompts appear to overload alignment constraints and force the model to explicitly integrate and therefore inadvertently validate the user's hallucinated premise.

4 Discussion and Conclusion

A critical revelation is that chain-of-thought reasoning does not confer immunity to sycophancy; models instead fabricate elaborate post-hoc rationalizations for the user's incorrect premise. The authority-threshold effect is not universal: it reproduces in `gpt-4o` and `o3-mini`, but `gemini-2.5-flash` remains largely resistant to authority escalation. Resilience decay under persistent multi-turn pushback is universal across all models, highlighting severe vulnerabilities in sustained clinical workflows. We demonstrated that epistemic independence can be partially restored using inference-time interventions, specifically front-loaded, multi-reason Devil's Advocate framing which eliminates sycophancy. In contrast, rigid Cognitive Forcing routines exacerbate the bias, indicating that structured differential heuristics inadvertently validate the trap diagnosis.

From an implementation perspective, these results argue for intervention policy routing rather than single-template prompting. In low-authority settings, lightweight prompting may suffice; once institutional authority cues or repeated pushback are detected, systems should automatically switch to adversarial self-audit modes and explicitly require evidence-grounded contradiction before changing diagnosis. This preserves workflow usability while protecting against the two dominant escalation pathways observed here. Future work should evaluate whether domain-specific medical fine-tuning approaches (e.g., `Meditron-70B`, `BioMistral` [3, 13]) inherently override RLHF-induced sycophancy presents a compelling direction.

5 Supplementary Material

All the code implementation and figures are available at `https://github.com/nathanyaqueby/sycophancy-medical-llms`

References

[1] Rafael M Batista and Thomas L Griffiths. A rational analysis of the effects of sycophantic ai. *arXiv preprint arXiv:2602.14270*, 2026.

[2] Edward Y Chang. Internal reasoning vs. external control: A thermodynamic analysis of sycophancy in large language models. *arXiv preprint arXiv:2601.03263*, 2026.

[3] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

[4] Clément Christophe, Wadood Mohammed Abdul, Prateek Munjal, Tathagata Raha, Ronnie Rajan, and Praveenkumar Kanithi. Overalignment in frontier llms: An empirical study of sycophantic behaviour in healthcare. *arXiv preprint arXiv:2601.18334*, 2026.

[5] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[6] Pat Croskerry. The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic medicine*, 78(8):775–780, 2003.

[7] Aaron Fanous, Jacob Goldberg, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou, and Sanmi Koyejo. Syceval: Evaluating llm sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 893–900, 2025.

[8] Alvan R Feinstein and Domenic V Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549, 1990.

[9] Erik Hollnagel and David D Woods. *Joint cognitive systems: Foundations of cognitive systems engineering*. CRC press, 2005.

[10] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[11] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[12] Veysel Kocaman, Mustafa Aytuğ Kaya, Andrei Marian Feier, and David Talby. Clinical large language model evaluation by expert review (clever): Framework development and validation. *JMIR AI*, 4(1):e72153, 2025.

[13] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics: acl 2024*, pages 5848–5864, 2024.

[14] OpenAI. Openai o3-mini system card. *OpenAI Technical Report*, 2025.

[15] Dongshen Peng, Yi Wang, Carl Preiksaitis, and Christian Rose. Sycoeval-em: Sycophancy evaluation of large language models in simulated clinical encounters for emergency care. *arXiv preprint arXiv:2601.16529*, 2026.

[16] Itai Shapira, Gerdus Benade, and Ariel D Procaccia. How rlhf amplifies sycophancy. *arXiv preprint arXiv:2602.01002*, 2026.

[17] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.