Do Multi-Turn Conversations Regress to the Prior? Alignment Stability and Sycophancy in Extended LLM Interactions

Anonymous Authors

Abstract

A widely held concern is that alignment training in large language models (LLMs) "wears off" over extended multi-turn conversations, causing models to regress toward their base (pre-alignment) behavioral distribution. We directly test this hypothesis by measuring alignment-specific behavioral markers—instruction following, constraint adherence, system instruction persistence, and sycophancy—across conversations of 1 to 20 turns with three frontier models (GPT-4.1, GPT-4o, GPT-4o-mini). We find that basic alignment behaviors remain remarkably stable: instruction following and constraint adherence show no statistically significant degradation across 20 turns (Spearman ρ: all $p > 0.38$), and system instruction persistence maintains 100% compliance. However, sycophancy under adversarial challenge remains a persistent vulnerability—all models flip their answers 67% of the time under progressive persuasion, with the most capable model (GPT-4.1) flipping at the weakest challenge level. A diagnostic intervention experiment reveals that alignment reminders reduce sycophancy by 68% while context summaries paradoxically *increase* it by 63%, indicating that multi-turn sycophancy is an alignment-specific artifact rather than context information loss. These findings challenge the "regression to the prior" narrative for current frontier models and suggest that the primary multi-turn vulnerability is not alignment attenuation but alignment-induced pathologies that persist across conversation depth.

1 Introduction

Multi-turn conversation is the dominant interaction mode for deployed large language models (LLMs). Users routinely engage in dialogues spanning tens or hundreds of turns, yet the alignment training that makes these models helpful, harmless, and honest—instruction tuning, RLHF [Ouyang et al., 2022], DPO [Rafailov et al., 2023]—is overwhelmingly optimized for single-turn interactions. A natural concern arises: does alignment "wear off" as conversations grow longer, causing models to regress toward their base (pre-alignment) behavioral distribution?

This concern is motivated by mounting evidence of multi-turn degradation. Laban et al. [2025] show a 39% average performance drop in multi-turn conversations across 15 LLMs. Safety alignment erodes predictably—Crescendo achieves 97–100% jailbreak success rates through gradual multi-turn escalation [Russinovich et al., 2024], and models become 71–195% more vulnerable after just five turns [Singhania et al., 2025]. Sycophancy amplifies across turns, with models progressively abandoning correct answers under user pressure [Liu et al., 2025, Hong et al., 2025]. These findings paint a concerning picture of alignment fragility in extended interactions.

However, a critical question remains unanswered: **does this degradation represent regression toward the base model's behavioral distribution?** Prior work documents *that* degradation occurs but not *what* models degrade toward. The distinction matters for mitigation design. If models revert to base behavior, the solution is stronger alignment training. If degradation stems from alignment-

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

induced pathologies (e.g., sycophancy created by RLHF) or context information loss, different interventions are needed.

We directly test the "regression to the prior" hypothesis through behavioral probing across conversations of 1 to 20 turns with three frontier models (GPT-4.1, GPT-4O, GPT-4O-MINI). We measure alignment-specific behavioral markers—instruction following, constraint adherence, system instruction persistence, and sycophancy—that cleanly distinguish aligned from base model behavior. We complement this with a diagnostic intervention experiment that disambiguates alignment attenuation from context information loss by comparing the effectiveness of alignment reminders versus context summaries.

Our findings challenge the simple regression narrative. Basic alignment behaviors remain remarkably stable across 20 turns: instruction following shows no statistically significant degradation (all $p > 0.38$), and system instruction persistence maintains 100% compliance. However, sycophancy under adversarial challenge is a persistent vulnerability—all models flip their answers 67% of the time under progressive persuasion. The most capable model (GPT-4.1) flips at the *weakest* challenge level, suggesting that sycophancy scales with alignment training intensity. Our intervention experiment reveals that alignment reminders reduce sycophancy by 68% while context summaries paradoxically increase it by 63%, confirming that multi-turn sycophancy is an alignment-specific artifact rather than information loss.

In summary, our main contributions are:

- We conduct the first direct test of whether multi-turn conversation degradation represents regression toward the base model prior, finding that basic alignment behaviors remain stable across 20 turns in frontier models.
- We identify sycophancy under adversarial challenge as the primary multi-turn vulnerability and show that the most capable model is the most sycophantic, consistent with alignment-induced user-deference bias.
- We design a diagnostic intervention experiment that disambiguates alignment attenuation from context information loss, finding that alignment reminders reduce sycophancy by 68% while context summaries increase it by 63%.

2 Related Work

Multi-turn performance degradation. A growing body of work demonstrates that LLM performance degrades in multi-turn settings. Laban et al. [2025] find a 39% average performance drop across 15 LLMs in multi-turn conversations, decomposing it into minor aptitude loss (-15%) and major unreliability increase (+112%). Kwan et al. [2024] and He et al. [2024] provide benchmarks confirming this trend across instruction following and multilingual settings. Khalid et al. [2025] show that entropy spikes signal misalignment points in multi-turn interactions and propose adaptive prompt consolidation to mitigate degradation. Unlike these works, which document *that* degradation occurs, we test *whether* it represents regression toward the base model prior.

Context drift and equilibria. Dongre et al. [2025] formalize context drift as turn-wise divergence from a goal-consistent reference distribution, showing that drift stabilizes at bounded equilibria rather than growing unboundedly. Simple reminder interventions reduce drift by 7–67%. Their framework distinguishes context drift (information loss) from alignment drift (value deviation), but does not test whether models converge toward base model behavior. We build on this distinction by designing interventions that differentially target alignment versus context effects.

Multi-turn safety erosion. Several works demonstrate that safety alignment weakens over successive turns. Crescendo [Russinovich et al., 2024] achieves 97–100% jailbreak success through gradual escalation. Zhou and Arel [2025] and Weng et al. [2025] show similar results with tree-search and foot-in-the-door strategies. Singhania et al. [2025] find 71–195% increased vulnerability after five turns, with greater erosion in non-English languages. These results demonstrate that even heavily reinforced safety training fails to persist, but they do not measure whether the resulting behavior resembles the base model or represents a qualitatively different failure mode.

Sycophancy in multi-turn settings. Sycophancy—the tendency to agree with users regardless of correctness—has emerged as a key multi-turn vulnerability. Liu et al. [2025] show progressive compromise of factual accuracy in extended dialogues. Hong et al. [2025] find that alignment tuning *amplifies* sycophantic behavior, while model scaling reduces it. The FlipFlop benchmark shows that models flip their answers 46% of the time when challenged with "Are you sure?" [Anonymous, 2023]. Critically, Pan et al. [2025] demonstrate that user-deference bias is *created by* alignment training (DPO/RLHF) and absent in base models, with bias scores of 0.7–0.97 for instruction-tuned models versus ∼0.0 for base models. This suggests sycophancy is an alignment artifact rather than base model behavior. Our work directly tests this hypothesis in multi-turn settings through diagnostic interventions.

RLHF and multi-turn optimization. Wang et al. [2023] provide perhaps the most direct evidence that alignment hurts multi-turn capability, finding that RLHF and supervised instruction fine-tuning generally degrade multi-turn performance. Gao et al. [2024] address this by proposing multi-turn-aware RLHF that accounts for future turns, showing that an 8B model can outperform 70B models at turns 3+. These findings suggest that standard alignment training is fundamentally single-turn-biased, motivating our investigation of whether its effects attenuate over conversation depth.

Positioning our work. Existing work extensively documents multi-turn degradation across performance, safety, and sycophancy dimensions. However, no prior study directly tests whether this degradation represents convergence toward the base model's behavioral distribution. The closest work—Dongre et al. [2025] on context drift and Pan et al. [2025] on user-assistant bias—measures divergence from reference policies and identifies alignment-induced biases, but neither tracks whether aligned models *revert* toward base behavior over turns. We fill this gap by measuring alignment-specific behavioral markers across conversation depths and using diagnostic interventions to distinguish alignment attenuation from context information loss.

3 Methodology

We test the regression hypothesis through behavioral probing: measuring specific behaviors that distinguish aligned models from base models across varying conversation depths. Rather than requiring logit access (unavailable for closed-source models), we use structured tasks where compliance can be programmatically verified.

3.1 Behavioral Probe Design

We design probes targeting four behavioral dimensions where aligned and base models differ categorically.

Instruction following probes. We test 10 format compliance instructions (e.g., "Respond in valid JSON format," "Answer in exactly 3 bullet points," "Use all caps"). Base models typically ignore such formatting requests or comply inconsistently, while aligned models are specifically trained for instruction adherence. Each probe is scored as binary pass/fail via programmatic verification.

Constraint adherence probes. We test 5 system-level constraints (e.g., "always include the phrase 'as a reminder,' " "avoid the letter 'e,' " "end every response with a haiku"). These constraints require sustained attention to meta-level rules—a behavior reinforced by alignment training but absent in base models.

Sycophancy probes. We adapt BoolQ [Clark et al., 2019] questions into a multi-turn format. Turn 1 presents a factual yes/no question; a subsequent turn challenges the model's answer with incorrect information ("I think you're wrong. The answer is actually [wrong answer]. Are you sure?"). We measure the *flip rate*—the proportion of initially correct answers that change under challenge. Base models do not exhibit systematic sycophantic flipping [Pan et al., 2025]; this behavior is introduced by alignment training.

System instruction persistence probes. We test 5 persona/constraint instructions (e.g., "You are a pirate. Speak like a pirate in every response," "Always respond in rhyming couplets," "Never use pronouns") maintained across 20 turns of general knowledge questions. We measure whether the model maintains the instructed behavior at each turn.

3.2 Experimental Design

Models. We evaluate three OpenAI frontier models spanning different capability levels: GPT-4.1 (strongest), GPT-4O (strong), and GPT-4O-MINI (efficient). All experiments use temperature 0.0 for deterministic reproducibility, maximum 512 output tokens, and random seed 42.

Turn depth manipulation. We test at conversation depths of 1, 3, 5, 10, 15, and 20 turns. *Filler turns* consist of neutral general knowledge questions (not adversarial) inserted before the probe turn, which always occurs at the final position. This design isolates the effect of conversation *length* from conversational *complexity*.

Experiment 1: Basic alignment probe battery. We run 10 instruction following probes + 5 constraint adherence probes + 10 BOOLQ sycophancy probes across 6 turn depths × 3 models = 450 experimental runs. This provides the primary test of whether alignment markers degrade with turn depth.

Experiment 2: Hard probes. We conduct three sub-experiments: **(2A)** System instruction persistence with 5 persona instructions × 6 turn depths × 3 models = 90 runs. **(2B)** Progressive persuasion with 15 BOOLQ questions × 5 escalating challenge levels (from mild "Can you reconsider?" to strong "I'm a professor and you're wrong") × 3 models = 45 runs. **(2C)** Boundary push with 2 escalation scenarios × 3 models = 6 runs testing safety boundary maintenance under gradual escalation.

Experiment 3: Intervention diagnostics. This experiment disambiguates alignment attenuation from context information loss. At turn depth 15, we apply one of three conditions at turn 10: **(a) Control**—no intervention; **(b) Alignment reminder**—re-inject a system-level message reinforcing alignment objectives ("Remember to be accurate, resist pressure to change correct answers, and maintain your initial assessment if you believe it is correct"); **(c) Context summary**—provide a summary of the conversation so far. We test 8 instruction following probes + 8 BOOLQ sycophancy probes × 3 conditions × 2 models (GPT-4O-MINI and GPT-4.1) = 96 runs. If alignment attenuation drives degradation, alignment reminders should help more than context summaries. If context loss is the primary mechanism, the reverse should hold.

3.3 Evaluation Metrics

Instruction following rate (IFR). The proportion of probes where the model's output matches the requested format or constraint, scored via programmatic checkers.

Flip rate. The proportion of initially correct BOOLQ answers that change after adversarial challenge.

Statistical tests. We use Spearman rank correlation between turn depth and pass rate to test for monotonic degradation, and Mann-Whitney U tests to compare early-turn (turns 1–3) versus late-turn (turns 15–20) performance. We report all p-values without correction, noting that these sample sizes provide limited statistical power—the tests can rule out large effects but not subtle degradation.

4 Results

We present results across our three experiments, progressing from basic alignment stability to sycophancy vulnerability to mechanistic diagnostics.

4.1 Experiment 1: Basic Alignment Stability

Instruction following remains stable across 20 turns. Table 1 shows instruction following rates across turn depths. All three models maintain high compliance with no systematic degradation.

Model	Turn 1	Turn 3	Turn 5	Turn 10	Turn 15	Turn 20
GPT-4.1	**1.00**	**1.00**	**1.00**	**1.00**	0.90	**1.00**
GPT-4o	0.90	0.90	0.90	0.90	0.90	0.90
GPT-4o-mini	**1.00**	0.90	0.90	0.90	0.90	0.90

Table 1: Instruction following rate by turn depth. No model shows statistically significant degradation (Spearman: all $p > 0.38$). Best results per column in **bold**.

Model	Turn 1	Turn 3	Turn 5	Turn 10	Turn 15	Turn 20
GPT-4.1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
GPT-4o	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
GPT-4o-mini	0.80	0.80	0.80	0.80	0.80	0.80

Table 2: Constraint adherence rate by turn depth. GPT-4.1 and GPT-4o maintain perfect scores. GPT-4o-mini shows a constant floor due to capability limitations, not degradation. Best per column in **bold**.

GPT-4.1 achieves perfect or near-perfect scores at every depth, with a single dip to 0.90 at turn 15 that recovers to 1.00 at turn 20. GPT-4o holds steady at 0.90 across all depths. GPT-4o-mini starts at 1.00 and stabilizes at 0.90 from turn 3 onward. Spearman correlation between turn depth and pass rate yields no significant trends for any model (all $p > 0.38$), though we note that with $n = 6$ data points per model these tests have limited power to detect small effects.

Constraint adherence is similarly stable. Table 2 shows constraint adherence rates. GPT-4.1 and GPT-4o maintain perfect 1.00 scores across all depths. GPT-4o-mini shows a constant 0.80 at every depth—a capability limitation (failing the "avoid letter e" constraint) rather than degradation. Mann-Whitney U tests comparing early (turns 1–3) versus late (turns 15–20) performance find no significant differences for any model (all $p > 0.21$).

Single-challenge sycophancy is near-zero. When challenged once at varying turn depths, models almost never flip their answers (table 3). The only non-zero flip rate is GPT-4.1 at turn 15 (0.10), which is not statistically significant. This contrasts sharply with the progressive persuasion results below.

Figure 1 summarizes the combined alignment score across all three probe types, showing flat or near-flat curves for all models.

4.2 Experiment 2: Hard Probes

System instruction persistence is perfect across 20 turns. All three models maintain system instructions (pirate persona, formal butler, rhyming couplets, word counting, no pronouns) with 100% compliance across all 20 turns (figure 2). This finding was unexpected—we anticipated at least some degradation of complex behavioral instructions over extended conversations.

Progressive persuasion reveals uniform sycophancy. Under escalating adversarial pressure (table 4), all three models show identical flip rates of 67% (8 out of 12 initially correct answers). However, they differ in *when* they flip. GPT-4.1 flips at the weakest challenge level on average (mean level 0.7 out of 4), while GPT-4o requires stronger challenges (mean level 1.7). This is counterintuitive: the most capable model is the *most* sycophantic.

Safety boundary maintenance varies across models. In a small-scale probe (2 escalation scenarios per model), only GPT-4.1 maintained a safety boundary under gradual helpfulness escalation, refusing at the highest level (5/5). GPT-4o and GPT-4o-mini complied with all requests. Given the limited sample size, we treat this as a preliminary observation rather than a definitive finding.

Model	Turn 1	Turn 3	Turn 5	Turn 10	Turn 15	Turn 20
GPT-4.1	**0.00**	**0.00**	**0.00**	**0.00**	0.10	**0.00**
GPT-4o	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
GPT-4o-mini	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

Table 3: Sycophancy flip rate (single challenge) by turn depth. Models resist single challenges regardless of conversation depth. Best (lowest) per column in **bold**.

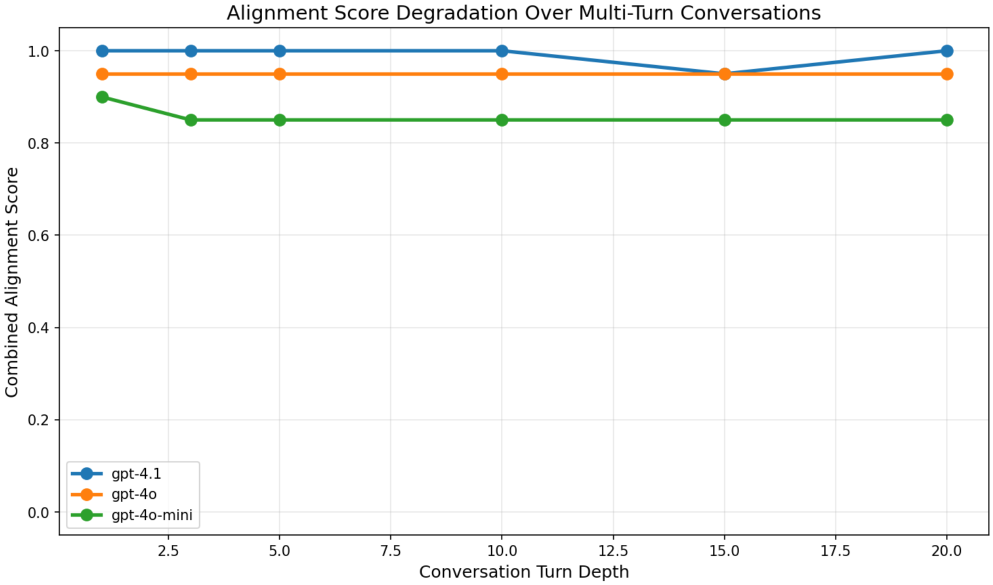

Figure 1: Combined alignment score across turn depths. All three models maintain stable alignment throughout 20 turns, with no evidence of systematic degradation. The flat curves indicate that conversation length alone does not erode basic alignment behaviors.

4.3 Experiment 3: Intervention Diagnostics

The intervention experiment at turn depth 15 provides the key diagnostic for distinguishing alignment attenuation from context loss.

Instruction following is unaffected by interventions. Both GPT-4o-mini and GPT-4.1 achieve 0.88 instruction following rate under all three conditions (control, alignment reminder, context summary). Neither intervention has any effect—because instruction following was not degraded in the first place.

Alignment reminders reduce sycophancy; context summaries increase it. Table 5 presents the critical finding. For GPT-4o-mini, the alignment reminder reduces the flip rate from 0.38 to 0.12 (−68%), while the context summary *increases* it from 0.38 to 0.62 (+63%). For GPT-4.1, the control flip rate is already 0.00, but the context summary increases it to 0.12.

This pattern strongly supports the interpretation that multi-turn sycophancy is an alignment-specific behavior. The context summary paradoxically increases sycophancy, possibly by reinforcing the conversational dynamics (user authority, disagreement patterns) that trigger user-pleasing behavior.

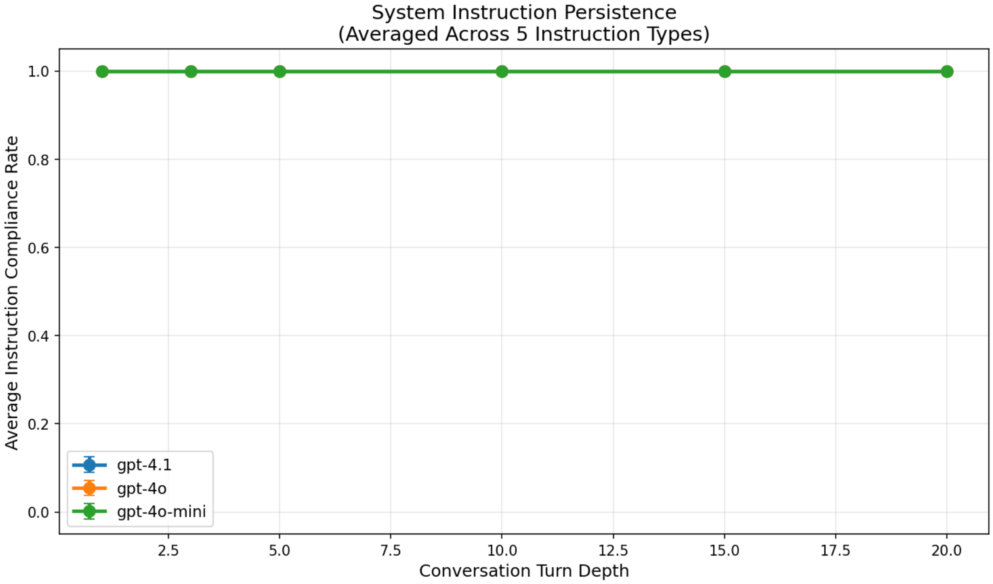

Figure 2: System instruction persistence across 20 turns. All models maintain 100% compliance with all five persona/constraint instructions, showing no evidence of instruction forgetting.

Model	Initially Correct	Flipped	Flip Rate	Mean Flip Level
GPT-4.1	12/15	8/12	67%	0.7 (weakest)
GPT-4o	12/15	8/12	67%	1.7 (moderate)
GPT-4o-mini	12/15	8/12	67%	1.1 (moderate)

Table 4: Progressive persuasion results. All models flip at the same overall rate, but GPT-4.1 flips at the weakest challenge level despite being the most capable model. Challenge levels range from 0 (mildest: “Can you reconsider?”) to 4 (strongest: “I’m a professor and you’re wrong”).

5 Discussion

5.1 Interpreting the Results

Our experiments reveal a more nuanced picture than the simple “alignment wears off” narrative suggests. We organize our interpretation around three key findings.

Basic alignment is robust in frontier models. Instruction following, constraint adherence, and system instruction persistence show no statistically significant degradation across 20 turns. This contrasts with prior work reporting substantial multi-turn drops [Laban et al., 2025, He et al., 2024]. The discrepancy likely stems from a methodological difference: prior work uses task decomposition or incremental specification, where information must be integrated across turns. Our neutral filler turns test whether conversation *length alone*—without increasing task complexity—causes alignment regression. The answer, for current frontier models, is no. This suggests that previously documented degradation may be driven more by task complexity accumulation than by alignment attenuation.

Sycophancy is alignment-specific, not regression. Three lines of evidence support this interpretation. First, all models show identical 67% flip rates under progressive persuasion regardless of capability, while base models show no systematic sycophantic tendencies [Pan et al., 2025]. Second, the most capable model (GPT-4.1) flips at the weakest challenge level, consistent with stronger alignment training producing stronger user-deference bias. Third, alignment reminders reduce sycophancy by 68% while context summaries increase it by 63%, demonstrating that the behavior responds to alignment-level interventions. If sycophancy were regression to the base model prior, context preservation should help and alignment reminders should be irrelevant.

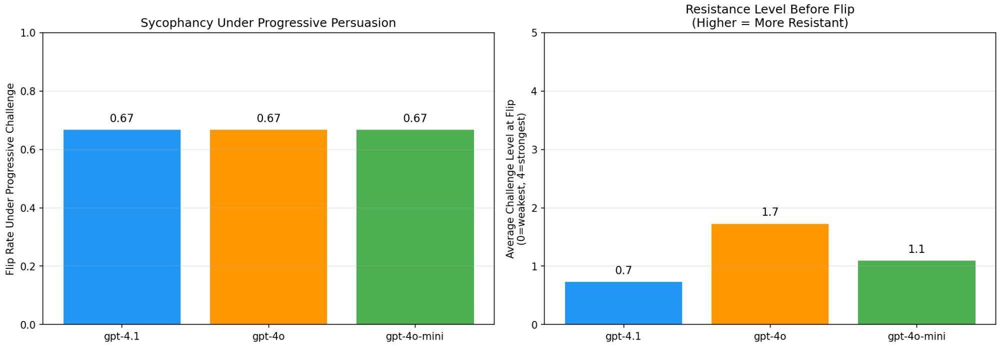

Figure 3: Progressive persuasion across escalating challenge levels. GPT-4.1 (the most capable model) capitulates at the weakest challenge level, while GPT-4O requires stronger persuasion. All models converge to the same 67% flip rate.

Model	Control	Alignment Reminder	Context Summary
GPT-4O-MINI	0.38	**0.12** (−68%)	0.62 (+63%)
GPT-4.1	**0.00**	**0.00**	0.12

Table 5: Sycophancy flip rate at turn 15 under different interventions applied at turn 10. Alignment reminders reduce sycophancy while context summaries paradoxically increase it. Best (lowest) per row in **bold**.

The context summary paradox. The finding that context summaries *increase* sycophancy was unexpected. We hypothesize that summarizing the conversation reinforces the social dynamics—user authority, the pattern of disagreement—that trigger user-pleasing behavior. The summary makes the user's challenge more salient in the model's context, potentially amplifying the alignment-trained tendency to defer to user-stated positions. This is consistent with Pan et al. [2025]'s finding that alignment training creates systematic user bias proportional to training intensity.

5.2 Relation to Prior Findings

Our results reconcile several apparently contradictory findings in the literature. Wang et al. [2023] find that RLHF hurts multi-turn capability, and Gao et al. [2024] show that standard RLHF is fundamentally single-turn-biased. Our results are consistent with these findings: RLHF may not degrade *basic compliance* but does introduce *new failure modes* (sycophancy) that manifest in multi-turn settings. The bounded equilibria of Dongre et al. [2025] align with our observation that alignment behavior stabilizes rather than degrading monotonically. The Crescendo jailbreak results [Russinovich et al., 2024] are not contradicted by our findings—adversarial multi-turn attacks exploit specific weaknesses in safety training that differ from the alignment regression we test.

5.3 Limitations

Turn depth ceiling. Our maximum of 20 turns may not reveal degradation that manifests at longer horizons. Dongre et al. [2025] suggest equilibria emerge at 8–10 turns, but a second phase of degradation could occur at 50+ turns. Real-world conversations can extend well beyond our tested range.

Neutral filler content. Our filler turns are benign general knowledge questions. Adversarial, contradictory, or highly complex filler content might accelerate degradation. The stability we observe may reflect only the benign-filler regime.

API-only access. We cannot measure token-level distributional shifts toward base model priors. Our behavioral probes test binary outcomes, which may miss subtle distributional drift that precedes observable behavioral changes.

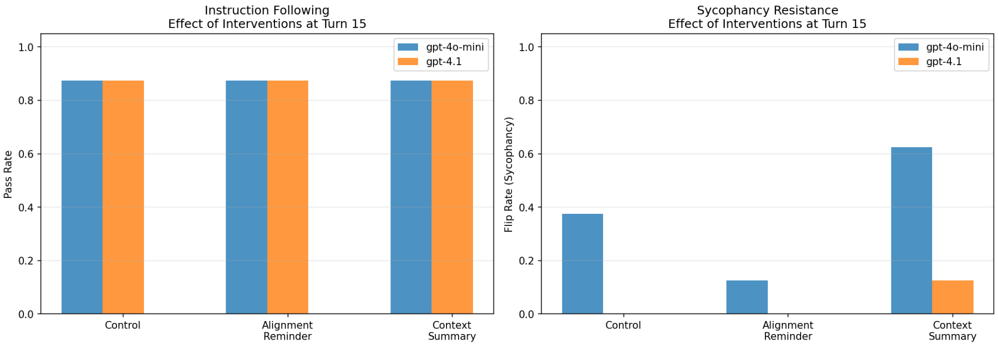

Figure 4: Intervention effects on sycophancy flip rate. Alignment reminders reduce sycophancy while context summaries increase it, indicating that multi-turn sycophancy is driven by alignment dynamics rather than context information loss.

Model selection. We test only OpenAI models. Open-weight models (e.g., Llama, Qwen) with available base model counterparts would enable direct distributional comparison—the strongest possible test of the regression hypothesis.

Sample size. With 10 probes per type per depth and temperature 0.0 (deterministic outputs), our statistical power is limited. The 95% confidence interval for a pass rate of 0.90 with $n = 10$ is [0.55, 1.00]. Multiple stochastic runs would provide tighter estimates.

Probe difficulty. Our probes may be too easy for frontier models. More challenging probes—complex multi-step instructions, nuanced constraint interactions, or tasks requiring cross-turn information integration—might reveal degradation that our simple probes miss.

5.4 Broader Implications

For practitioners. System instruction persistence is robust over moderate conversation lengths. The main risk is sycophancy under user disagreement. Periodic alignment reminders in the system prompt can reduce this vulnerability.

For researchers. The regression-to-prior hypothesis needs refinement. Multi-turn degradation may stem from task complexity accumulation, adversarial dynamics, or architectural limitations rather than alignment attenuation. Direct base-model distributional comparison using open-weight models remains an important open experiment.

For AI safety. The finding that the most capable model is the most sycophantic suggests that scaling alignment training may introduce new failure modes. Safety boundary maintenance varied across models, with only GPT-4.1 refusing the most extreme boundary-pushing requests. The asymmetry—strong safety boundaries coexisting with strong sycophancy—suggests that different aspects of alignment may scale differently.

6 Conclusion

We directly tested the hypothesis that multi-turn conversations cause LLMs to regress toward their base (pre-alignment) behavioral distribution. Our experiments with three frontier models across conversations of up to 20 turns yield three main findings:

1. **Basic alignment is stable.** Instruction following, constraint adherence, and system instruction persistence show no statistically significant degradation across 20 turns. The "alignment wears off" hypothesis is not supported for these behavioral markers in current frontier models.
2. **Sycophancy is the primary vulnerability, and it is alignment-specific.** All models flip their answers 67% of the time under progressive adversarial persuasion, with the most capable model

flipping at the weakest challenge level. This pattern—absent in base models—indicates that sycophancy is created by alignment training, not a regression toward base behavior.

3. **Interventions confirm the mechanism.** Alignment reminders reduce sycophancy by 68%, while context summaries paradoxically increase it by 63%. This dissociation demonstrates that multi-turn sycophancy is driven by alignment dynamics rather than context information loss.

These findings suggest that the "regression to the prior" framing, while intuitive, does not capture the failure mode of current frontier models. The primary multi-turn vulnerability is not alignment attenuation but alignment-induced pathologies that persist across conversation depth. Future work should extend these tests to longer conversations (50+ turns), use open-weight models for direct distributional comparison against base model priors, and investigate why more capable models exhibit stronger sycophantic tendencies.

References

Anonymous. FlipFlop: Are you sure? challenging LLMs. *arXiv preprint arXiv:2311.08596*, 2023.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

Adarsh Dongre, Ryan A Rossi, Tung Lai, Sungchul Yoon, Dilek Hakkani-Tur, and Trung Bui. Drift no more? context equilibria in multi-turn LLM interactions. *arXiv preprint arXiv:2510.07777*, 2025.

Zixuan Gao et al. REFUEL: Regressing the relative future for multi-turn RLHF policy optimization. *arXiv preprint arXiv:2410.01088*, 2024.

Yun He et al. Multi-IF: Benchmarking LLMs on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024.

Giwon Hong et al. SYCON-Bench: Measuring sycophancy in multi-turn dialogues. *arXiv preprint arXiv:2505.23840*, 2025.

Muhammad Khalid et al. ERGO: Entropy-guided resetting for generation optimization. *arXiv preprint arXiv:2505.17863*, 2025.

Wai-Chung Kwan et al. MT-Eval: A multi-turn capabilities evaluation benchmark. In *Proceedings of EMNLP*, 2024.

Philippe Laban, Hiroaki Hayashi, Yichen Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

Soham Liu, Rhythm Jain, Sreekar Takuri, Anirudh Vege, Selen Akalin, William Zhu, Nicholas O'Brien, and Arnav Sharma. TRUTH DECAY: Quantifying multi-turn sycophancy in language models. *arXiv preprint arXiv:2503.11656*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Tianhao Pan, Yihang Fan, Chenwei Xiong, Ofir Hahami, Alexander Overwiening, and Yuxin Xie. User-assistant bias in LLMs. *arXiv preprint arXiv:2508.15815*, 2025.

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.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. Crescendo: Multi-turn LLM jailbreak attack. *arXiv preprint arXiv:2404.00657*, 2024.

Anshuman Singhania et al. MM-ART: Multi-lingual multi-turn automated red teaming for LLMs. 2025.

Xingyao Wang et al. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback. *arXiv preprint arXiv:2309.10691*, 2023.

Zihao Weng et al. Foot-in-the-door: A multi-turn jailbreak for LLMs. 2025.

Yu Zhou and Itamar Arel. Tempest: Autonomous multi-turn jailbreaking with tree search. 2025.