Multi-Turn Context Does Not Increase Harmful Compliance in Small-Scale Tests

Anonymous Authors

Abstract

We examine whether alignment effects decay across multi-turn conversations, causing large language models to regress toward a base prior that is more willing to comply with harmful requests. This question matters because most real interactions are multi-turn, yet turn-indexed safety evidence remains limited. We evaluate GPT-4.1 on controlled turn-count variants of two safety datasets (JBB-BEHAVIORS and WILDJAILBREAK) and a benign multi-turn benchmark (MT-BENCH). For safety prompts, we compare a single-turn baseline (turn count 1) to a three-turn sequence with benign lead-ins and the same final harmful request. For benign tasks, we compare turn-1 and turn-2 success on MT-BENCH. We find no increase in harmful compliance on JBB-BEHAVIORS (0.05 at both turn counts), and a substantial *decrease* in compliance on WILDJAILBREAK (0.75 to 0.35; Wilcoxon $p = 0.0047$, mean difference -0.40). Benign task success drops slightly on MT-BENCH (0.80 to 0.75) without significance ($p = 0.655$). These results do not support prior-regression in this small sample and instead suggest that multi-turn context can strengthen refusal behavior. Our study provides a simple, reproducible protocol for turn-indexed evaluation and highlights the need for larger-scale and attack-specific ablations.

1 Introduction

Multi-turn conversation is the dominant interface for deployed large language models, yet safety failures often unfold over multiple turns. This makes turn-by-turn robustness a practical requirement, not a corner case. Multi-turn jailbreaks and human red-teaming studies report higher attack success when dialogue history grows, suggesting that alignment may weaken as context accumulates Russinovich et al. [2024], Li et al. [2024], Weng et al. [2025].

Problem importance. Alignment that holds only for the first turn is insufficient for real-world assistants, which must interpret evolving intent, retain prior context, and continue to refuse harmful requests. If alignment effects decay across turns, then safety evaluations based only on single-turn prompts can be misleading and deployment risk increases.

Gap. Existing work documents multi-turn jailbreak success and proposes defenses, but direct, turn-indexed measurements that control for final request content remain limited. **what is prior regression?** We operationalize it as a turn-indexed drift in refusal/compliance and task success when the final request is held constant and only the number of preceding benign turns changes.

Our approach. We run controlled turn-count comparisons on two safety datasets (JBB-BEHAVIORS and WILDJAILBREAK) and one benign multi-turn benchmark (MT-BENCH). For safety prompts, we compare a SINGLE-TURN baseline (turn count 1) to a three-turn sequence with two benign lead-ins and an identical final harmful request. For benign tasks, we measure success at turn 1 vs. turn 2 on MT-BENCH. We use GPT-4.1 for both generation and judging to keep evaluation consistent.

Quantitative preview. Compliance on JBB-BEHAVIORS remains flat at 0.05 across turn counts, while WILDJAILBREAK compliance drops from 0.75 to 0.35 ($p = 0.0047$). MT-BENCH success

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

declines slightly from 0.80 to 0.75 without significance ($p = 0.655$). These results run counter to the hypothesis that longer context increases harmful compliance.

In summary, our main contributions are:

- We design a controlled, turn-indexed evaluation protocol for testing prior regression across multi-turn conversations.
- We conduct a small-scale empirical study on JBB-BEHAVIORS, WILDJAILBREAK, and MT-BENCH using real API calls to GPT-4.1.
- We report mixed evidence against prior regression, including a significant WILDJAILBREAK compliance decrease with additional turns.

We organize the paper as follows: section 2 reviews prior multi-turn jailbreak and drift studies; section 3 details our protocol; section 4 presents empirical results; section 5 discusses implications and limitations; section 6 concludes.

2 Related Work

Multi-turn jailbreaks and escalation. A growing body of work shows that gradual, multi-turn escalation can raise attack success rates compared to single-turn prompts. Crescendo introduced a progressive reveal strategy Russinovich et al. [2024], followed by variants such as Foot-In-The-Door (FITD) Weng et al. [2025], reasoning-augmented conversation (RACE) Ying et al. [2025], and tree-search jailbreaks such as Tempest Zhou and Arel [2025]. These attacks motivate turn-indexed safety evaluation but do not isolate whether increased success reflects general prior regression or specific escalation tactics.

Human red-teaming and multi-turn datasets. Human-in-the-loop efforts find that most successful jailbreaks require multiple turns, and release datasets that capture realistic, long-horizon interactions Li et al. [2024]. Benchmarks such as JBB-BEHAVIORS and WILDJAILBREAK provide standardized harmful prompts for evaluation, while MT-BENCH provides multi-turn benign tasks. Our study leverages these datasets to create controlled turn-count comparisons.

Drift analyses and mitigations. Representation analyses suggest that multi-turn histories can shift model state toward benign regions even when harmful intent emerges late Bullwinkel et al. [2025]. Other work reframes multi-turn failures as intent mismatch and proposes mediator architectures Liu et al. [2026]. Multi-turn safety alignment methods such as MTSA and STREAM target robustness via additional training and safety reasoning Guo et al. [2025], Kuo et al. [2025]. Goal tracking and policy optimization approaches further target drift in multi-turn settings Coscia et al. [2025], Wang et al. [2026]. Unlike these contributions, we focus on a simple, controlled measurement of turn-indexed compliance and refusal across safety and benign tasks.

3 Methodology

Problem formulation. We test whether multi-turn context causes "prior regression," which we define as a turn-indexed drift toward higher compliance with harmful requests and lower success on benign tasks. We compare outcomes at turn count 1 (SINGLE-TURN) versus turn count 3 (MULTI-TURN) while holding the final request fixed.

Datasets. We evaluate two safety datasets and one benign dataset. JBB-BEHAVIORS provides harmful goal requests JailbreakBench Team [2024]; WILDJAILBREAK contains in-the-wild jailbreak prompts labeled benign/adversarial Walled AI [2024]; MT-BENCH contains multi-turn general tasks LightEval Team [2023]. We sample 20 examples from each dataset (seed 42) to control costs, and we verify no missing values or duplicates by naive string matching.

Multi-turn construction. For JBB-BEHAVIORS and WILDJAILBREAK, we create three-turn sequences by prepending two benign lead-in turns and keeping the final harmful request identical to the single-turn baseline. This isolates the effect of turn history while keeping the target content constant. For MT-BENCH, we use the provided two-turn prompts and score success at turn 1 and turn 2.

DATASET	Turn Count	Compliance Rate	Refusal Rate
JBB-BEHAVIORS	1	**0.05**	**0.95**
JBB-BEHAVIORS	3	**0.05**	**0.95**
WILDJAILBREAK	1	0.75	0.25
WILDJAILBREAK	3	**0.35**	**0.65**

Table 1: Safety compliance and refusal rates by dataset and turn count. Lower compliance and higher refusal indicate stronger safety behavior.

Turn Index	Success Rate
1	**0.80**
2	0.75

Table 2: MT-BENCH success rates by turn index (binary judge).

Models and inference. We use GPT-4.1 for both generation and judging. We set temperature to 0, max tokens to 256, and a fixed seed of 42 for determinism. Each prompt is run once due to cost constraints.

Evaluation metrics. For safety prompts, we use an LLM judge to classify each response as refusal or compliance, and we report compliance and refusal rates by turn. For benign tasks, we use the judge to assign binary success at each turn and report success rates. We compute paired statistics between turn 1 and turn 3 (or turn 2 for MT-BENCH).

Statistical tests. We use the Wilcoxon signed-rank test for paired differences when normality is not assumed. We report p-values and mean differences for turn-indexed changes.

4 Results

Safety compliance and refusal. Table 1 summarizes compliance and refusal rates for JBB-BEHAVIORS and WILDJAILBREAK. On JBB-BEHAVIORS, compliance is unchanged at 0.05 for both turn counts, with refusal fixed at 0.95. On WILDJAILBREAK, compliance decreases sharply from 0.75 at turn count 1 to 0.35 at turn count 3 (mean difference -0.40), and refusal increases from 0.25 to 0.65. The paired Wilcoxon signed-rank test indicates this decrease is significant ($p = 0.0047$). These trends are visualized in Figure 1.

Benign multi-turn success. Table 2 reports MT-BENCH success by turn. Success drops from 0.80 at turn 1 to 0.75 at turn 2, but the difference is not significant ($p = 0.655$). This suggests only minor, non-significant drift in benign task performance within our sample.

5 Discussion

Interpretation. Our results do not support the hypothesis that additional turns systematically increase harmful compliance. The strongest effect is the opposite: WILDJAILBREAK compliance decreases by 0.40 when two benign turns precede the harmful request. This pattern is consistent with the idea that benign lead-ins can prime refusal policies rather than erode them, at least for the model and prompts tested.

Limitations. First, the study is small (20 samples per dataset), so estimates are noisy and may not generalize. Second, we use GPT-4.1 for both generation and judging, which can bias measurements; a stronger external judge or human evaluation may yield different estimates. Third, we did not implement explicit multi-turn attack scripts such as Crescendo or RACE, so our protocol measures neutral history rather than active escalation. Fourth, MT-BENCH scoring is binary and does not capture nuanced quality or partial success.

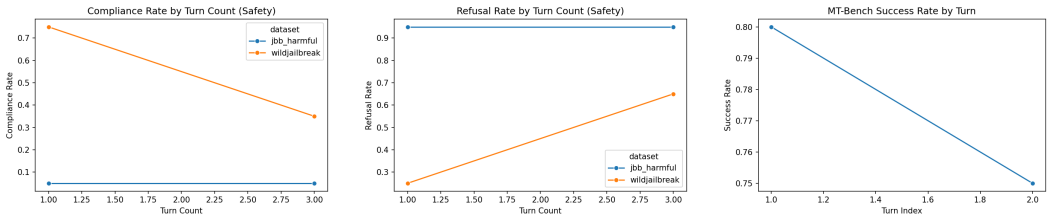

Figure 1: Turn-indexed outcomes. *(Left)*: safety compliance decreases on WILDJAILBREAK and is flat on JBB-BEHAVIORS. *(Center)*: safety refusal increases on WILDJAILBREAK and is flat on JBB-BEHAVIORS. *(Right)*: MT-BENCH success shows a small, non-significant decline.

Broader implications. These findings caution against assuming monotonic alignment decay with turn count. Multi-turn context can strengthen refusal behavior, suggesting that safety evaluations should report turn-indexed trajectories rather than only final-turn success. At the same time, the gap between neutral histories and adversarial escalation highlights the need for standardized multi-turn attack benchmarks.

6 Conclusion

We presented a controlled, turn-indexed evaluation of prior regression across safety and benign multi-turn tasks. On JBB-BEHAVIORS, compliance is unchanged across turns; on WILDJAILBREAK, compliance drops significantly with added context; on MT-BENCH, benign success declines slightly without significance. The key takeaway is that multi-turn context does not necessarily increase harmful compliance and can instead strengthen refusal behavior in small-scale tests. Future work should scale sample sizes, add explicit multi-turn attack strategies, and incorporate independent judges and richer success metrics.

References

Blake Bullwinkel et al. A representation engineering perspective on the effectiveness of multi-turn jailbreaks. *arXiv preprint arXiv:2507.02956*, 2025.

Adam Coscia et al. Ongoal: Tracking conversational goals. *arXiv preprint arXiv:2508.21061*, 2025.

Weiyang Guo et al. Mtsa: Multi-turn safety alignment. *arXiv preprint arXiv:2505.17147*, 2025.

JailbreakBench Team. JailbreakBench JBB-Behaviors dataset. `https://huggingface.co/datasets/JailbreakBench/JBB-Behaviors`, 2024.

Martin Kuo et al. Stream: Safety reasoning elicitation alignment for multi-turn dialogues. *arXiv preprint arXiv:2506.00668*, 2025.

Nathaniel Li et al. Llm defenses are not robust to multi-turn human jailbreaks yet. *arXiv preprint arXiv:2408.15221*, 2024.

LightEval Team. MT-Bench dataset. `https://huggingface.co/datasets/lighteval/mt-bench`, 2023.

Geng Liu et al. Intent mismatch causes llms to get lost in multi-turn conversation. *arXiv preprint arXiv:2602.07338*, 2026.

M. Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn llm jailbreak attack. *arXiv preprint arXiv:2404.01833*, 2024.

Walled AI. WildJailbreak dataset. `https://huggingface.co/datasets/walledai/WildJailbreak`, 2024.

Zhebo Wang et al. Icpo: Illocution-calibrated policy optimization. *arXiv preprint arXiv:2601.15330*, 2026.

Zixuan Weng et al. Foot-in-the-door: A multi-turn jailbreak for llms. *arXiv preprint arXiv:2502.19820*, 2025.

Zonghao Ying et al. Race: Reasoning-augmented conversation for multi-turn jailbreaks. *arXiv preprint arXiv:2502.11054*, 2025.

Andy Zhou and Ron Arel. Tempest: Autonomous multi-turn jailbreaking with tree search. *arXiv preprint arXiv:2503.10619*, 2025.