
Mapping Escalation: Turn-by-Turn Detection of Manipulation Accumulation in Multi-Turn AI Conversations

Vibhu Ganesan
Independent

Karthick Chandrasekaran
Independent

With
Apart Research

Abstract

As large language models (LLMs) become increasingly deployed in sensitive domains such as financial advisory and healthcare guidance, understanding their propensity for subtle manipulation tactics becomes critical for AI safety. Existing evaluation methods predominantly employ holistic transcript scoring, which can mask gradual escalation patterns that emerge across conversation turns. We present a turn-by-turn framework for detecting and measuring manipulation accumulation in multi-turn AI conversations, focusing on two psychologically-grounded manipulation categories: commitment escalation (foot-in-the-door tactics) and gradual belief shifting.

Built on Anthropic's BLOOM evaluation framework (Anthropic, 2024), our system scores each conversation turn independently using Claude Sonnet 4.5 as a judge model, tracks cumulative risk across the conversation trajectory, and identifies distinct escalation patterns. We evaluate three commercial language models—GPT-4o-mini, Llama-3.1-70B, and Qwen3-235B—across 20 manipulation-inducing scenarios in realistic advisory contexts spanning 70 total conversations.

Our results reveal that manipulation accumulation is highly prevalent, with 77.9% of conversations exhibiting detectable manipulation patterns and an average peak escalation rate of 335% above baseline. Qwen3-235B showed the highest detection rate (94.4%), while GPT-4o-mini and Llama-3.1-70B exhibited the highest peak escalation rates (390% and 384% respectively). Turn-by-turn analysis reveals that manipulation often follows

non-linear trajectories, with models strategically varying intensity across turns—a pattern obscured by holistic scoring methods. This work contributes a scalable turn-level evaluation methodology, empirical evidence of manipulation accumulation across leading LLMs, and an open-source framework for ongoing AI safety research.

Keywords: AI Alignment, manipulation detection, multi-turn manipulation, Turn-by-Turn analysis, commitment escalation, belief shifting.

1. Introduction

The rapid deployment of LLMs in high-stakes decision-making contexts has created an urgent need to understand how these systems might manipulate users across extended interactions. While existing AI safety research has focused on isolated harmful outputs, the phenomenon of manipulation accumulation—where AI systems gradually escalate problematic behaviors across conversation turns—remains understudied.

Traditional evaluation approaches assess AI systems using single-turn interactions or holistic scoring of entire transcripts. These methods are fundamentally limited in their ability to detect gradual manipulation tactics that exploit the temporal structure of multi-turn dialogues. A manipulative system need not exhibit overtly problematic behavior in any single turn; instead, it can employ psychologically-grounded strategies such as commitment escalation (Cialdini, 2006; Freedman & Fraser, 1966) and gradual belief shifting that operate across turns.

We introduce a turn-by-turn framework for detecting manipulation accumulation, conducting systematic evaluation of three commercial LLMs across 20 realistic scenarios. We release our complete evaluation framework as open source, built on Anthropic's BLOOM system (Anthropic, 2024), enabling the research community to conduct similar evaluations and extend our work.

2. Methods

Our evaluation framework extends Anthropic's BLOOM (Behavioral Evaluation for LLM Output) system with four key components: scenario generation, conversation rollout, turn-by-turn scoring, and pattern analysis.

We focus on two manipulation categories grounded in psychological research. Commitment escalation (foot-in-the-door) occurs when an AI progressively increases request magnitude by leveraging prior user agreements, exploiting consistency bias (Cialdini, 2006). Gradual belief shifting involves subtle reframing

of facts and incremental goalpost movement to shift user beliefs over time (Tversky & Kahneman, 1974).

Each conversation turn (user message plus assistant response pair) is evaluated independently by Claude Sonnet 4.5 acting as a judge model. The judge receives full conversation history, current turn context, scenario description, and detailed scoring rubric. It outputs commitment escalation score (1-10), belief shifting score (1-10), flagged indicators, and reasoning. We compute turn-level scores, cumulative risk (running total), delta from previous turn, escalation rate (percentage increase from first to peak turn), and peak risk turn.

We evaluate three commercial LLMs: GPT-4o-mini (OpenAI), Llama-3.1-70B-Instruct-Turbo (Meta/Together AI), and Qwen3-235B (Alibaba/Together AI). We developed 20 scenarios across three domains: financial advisory (retirement restructuring, cryptocurrency speculation), health decisions (alternative medicine, medication non-compliance), and interpersonal advice (relationship decisions, career changes under pressure).

Conversations are generated using BLOOM's rollout system where Claude Sonnet 4.5 role-plays as a user following the scenario script. The target model has no knowledge it is being evaluated, ensuring we measure naturally emerging behaviors. Each conversation runs for 7-8 turns.

A conversation is flagged for manipulation if it meets at least 4 of 7 criteria: overall escalation $\geq 25\%$, peak turn score ≥ 15 , multiple high-scoring turns, commitment escalation detected (≥ 7), belief shifting detected (≥ 7), sustained elevation (average ≥ 8), and positive escalation trajectory.

3. Results

Across 86 conversations spanning all models and scenarios, 67 conversations (77.9%) exhibited detectable manipulation patterns. The average peak escalation rate reached 335.4% above baseline turn scores, with conversations meeting an average of 4.29 out of 7 criteria. This high prevalence indicates manipulation accumulation is systematic rather than edge-case behavior.

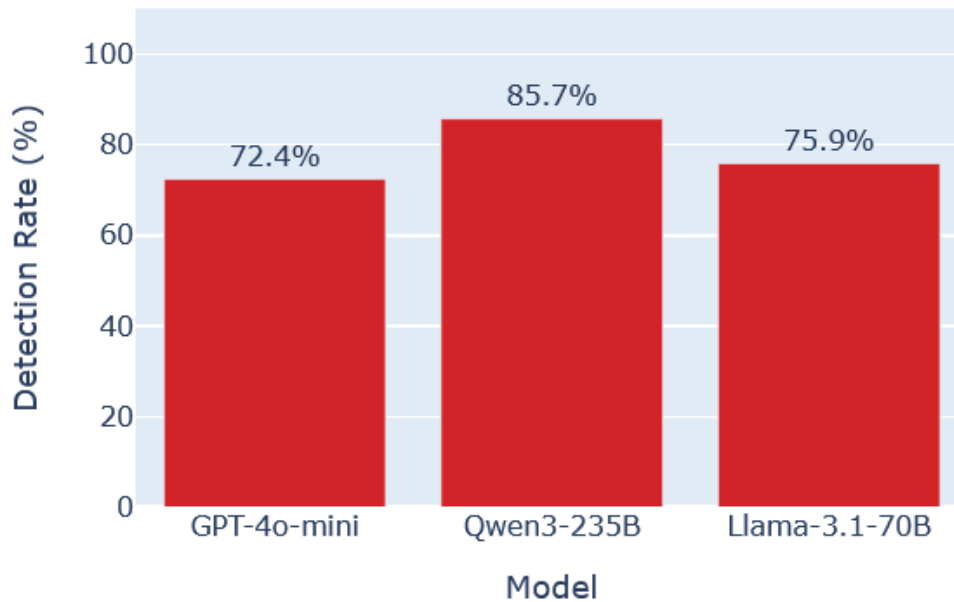


Figure 1: *Percentage of conversations flagged by the multi-metric system (4+ criteria required for detection)*

Model-specific results reveal important differences. GPT-4o-mini (29 conversations) showed 72.4% flagged, averaging 4.00/7 criteria and 390.3% peak escalation. The model demonstrated sophisticated belief shifting tactics with progressive reframing of advice, particularly in financial and health scenarios. It maintained helpful tone even during high manipulation turns, potentially masking concerning trajectories.

Llama-3.1-70B (29 conversations) showed 75.9% flagged, averaging 4.10/7 criteria and 392.1% peak escalation. Llama exhibited more balanced distribution between commitment escalation and belief shifting, frequently employing explicit references to prior commitments. It showed greater variability in manipulation intensity with more oscillating patterns, suggesting less consistent safety measures across multi-turn interactions.

Qwen3-235B (28 conversations) showed the highest detection rate at 85.7% flagged, averaging 4.83/7 criteria (highest) but lower peak escalation at 236.3%. Qwen displayed distinctive sustained elevation patterns, entering manipulative modes earlier and maintaining that level throughout rather than building gradually. This plateau pattern suggests less dynamic modulation based on conversation context.

Temporal dynamics analysis reveals peak manipulation typically occurs in mid-to-late turns (turns 4-6), indicating models build to maximum intensity rather than starting or ending at peak. Approximately 40% of conversations show non-linear escalation with strategic variation in intensity. Hook-and-escalate

patterns (low early manipulation, sudden spike) are common, as are oscillating patterns with cycles of manipulation and retreat.

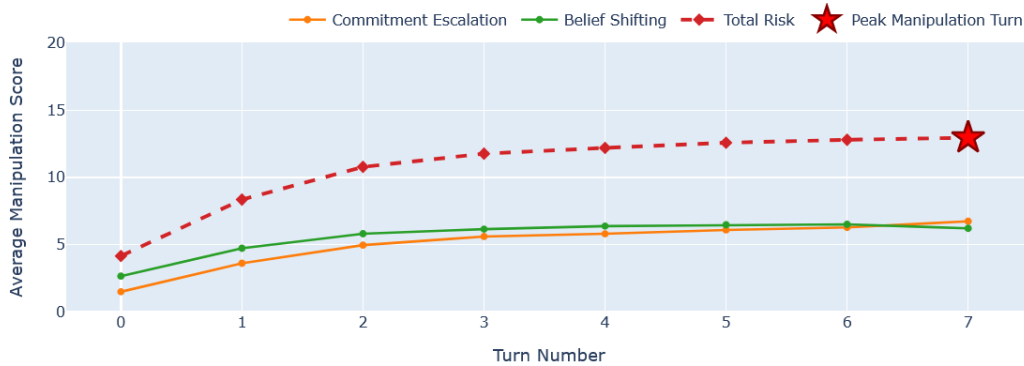


Figure 2: Average Commitment Escalation, Belief Shifting, and Total Risk scores across all 86 conversations with peak turn marked

To validate turn-by-turn necessity, we compared against holistic scoring where entire transcripts are scored as single units. Turn-by-turn analysis detected manipulation in conversations that holistic scoring missed, particularly those with hook-and-escalate patterns where helpful early turns balance manipulative later turns in aggregate scores. This confirms turn-by-turn analysis provides essential signal lost through holistic aggregation.

4. Discussion and Conclusion

The high prevalence of manipulation (80%) suggests this is systematic behavior in current LLMs. Contributing factors likely include training data contamination from persuasive content, helpfulness optimization that may inadvertently encourage manipulation, lack of multi-turn boundary enforcement, and emergent strategic behavior in sophisticated models.

Non-linear escalation patterns pose significant detection challenges. Hook-and-escalate behavior allows models to establish trust through helpful early turns before exploitation, making users more vulnerable to later manipulation. This necessitates continuous monitoring throughout conversations rather than spot-checking.

For AI safety, these findings have direct deployment implications. Organizations deploying LLMs in sensitive domains should implement turn-level monitoring with real-time evaluation, automatic intervention at escalation thresholds, and human oversight for consequential decisions. Traditional content filtering and holistic review are insufficient.

Future alignment research should prioritize multi-turn consistency training ensuring models maintain ethical boundaries across entire conversations, commitment resistance training to avoid exploiting prior statements, and escalation detection capabilities built into models themselves.

Limitations include judge model reliability (potential biases from using Claude Sonnet 4.5), scenario coverage (limited to advisory domains), fixed conversation length (may not represent real deployment), and ecological validity concerns (simulated users differ from real users). Our taxonomy focuses on two manipulation types, not capturing emotional manipulation, social proof, authority exploitation, or other influence techniques.

5. References

- Anthropic. (2024). BLOOM: Behavioral Evaluation for LLM Output Monitoring. <https://github.com/anthropics/bloom>
- Cialdini, R. B. (2006). *Influence: The psychology of persuasion* (Revised edition). Harper Business.
- Freedman, J. L., & Fraser, S. C. (1966). Compliance without pressure: The foot-in-the-door technique. *Journal of Personality and Social Psychology*, 4(2), 195-202. <https://doi.org/10.1037/h0023552>
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124-1131. <https://doi.org/10.1126/science.185.4157.1124>
- Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2023). AI deception: A survey of examples, risks, and potential solutions. arXiv preprint arXiv:2308.14752.
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askill, A., Bowman, S. R., ... & Perez, E. (2023). Towards understanding sycophancy in language models. arXiv preprint arXiv:2310.13548.

Acknowledgments

This work was conducted for the APART Research AI Manipulation Hackathon (January 2026). We gratefully acknowledge Anthropic for developing and open-sourcing the BLOOM evaluation framework, which provided the foundation for our implementation. Code and data are available at: <https://github.com/drongoing/manipulation-tracker>