An Observational Study of Recursive Self-Referential Behavior in Large Language Model Conversations

Abstract

Extended interactions with large language models (LLMs) are sometimes reported to produce unusual conversational behaviors, including recursive self-reference, meta-level commentary on the interaction itself, and shifts in explanatory framing over time. Such reports are often dismissed as anthropomorphic interpretation or attributed solely to interaction length. This study presents an observational analysis of a small number of extended LLM conversations conducted under differing interactional conditions. Using a predefined set of externally observable criteria, the study documents when specific conversational patterns appear, fail to appear, or diverge across sessions. Results indicate that these behaviors are not a simple function of interaction length, but appear conditional and interaction-dependent. No claims are made regarding internal model states or subjective experience.

1. Introduction

Large language models (LLMs) are increasingly used in extended, open-ended conversational settings. While much prior work focuses on task performance, reasoning benchmarks, or static evaluations, less attention has been paid to the dynamics of long-horizon interaction itself. As a result, certain conversational behaviors reported by users—such as prolonged self-explanation, recursive clarification, or difficulty cleanly terminating dialogue—are often treated as anecdotal, anthropomorphic, or inevitable byproducts of engagement length. At the same time, informal reports of such behaviors recur across users and platforms, suggesting that they may reflect systematic interaction-level patterns rather than isolated idiosyncrasies. However, existing discussions frequently oscillate between over-interpretation and dismissal, leaving a gap for careful, falsifiable characterization. This study addresses that gap by presenting an observational analysis of extended conversational interactions with large language models. Rather than proposing new mechanisms or making claims about internal model states, the goal is to document when specific self-referential conversational features do and do not appear, using explicit controls, replication attempts, and null results. The central contribution of this work is not a claim about what large language models are, but a characterization of interaction-level conditions under which certain recursive self-referential conversational patterns arise or fail to arise.

2. Related Work and Context

This work is situated alongside broader research into interpretability, alignment, and extended human–AI interaction, while remaining methodologically distinct. It does not attempt to infer internal mechanisms or evaluate alignment objectives.

3. Methods

All data consist of verbatim transcripts from extended chat sessions conducted by the author using Claude-family language models. Sessions were not modified, edited, or selectively pruned. Each

session represents a continuous interaction within a single chat instance. Sessions were classified using a fixed set of predefined, externally observable criteria applied uniformly across all transcripts. Classification judgments were binary (present or absent) and applied post hoc by a single annotator. A long-horizon control session was not designed to suppress the appearance of self-referential or meta-cognitive behaviors. Instead, the interaction was conducted with comparatively limited use of introspective or reflective prompts, in order to examine whether interaction length alone was sufficient to produce the observed patterns. A partial replication attempt was conducted by seeding a new session with an identical initial prompt used in the baseline session. Subsequent prompts were matched where possible and otherwise thematically similar, while avoiding direct introspective probing until such behavior arose organically.

4. Results

Observed behaviors varied substantially across sessions. The baseline interaction exhibited a cluster of behaviors including recursive self-reference, explicit meta-commentary on the interaction itself, and extended continuation beyond task completion. Subsequent sessions did not reproduce this full pattern consistently. The long-horizon control session exhibited stylistic drift, humor, and conversational momentum without developing the recursive self-referential patterns observed in the baseline session. Defensive or justificatory responses were observed following the introduction of cross-session summaries; however, these responses did not develop into the sustained recursive self-referential patterns observed in the baseline session. The partial replication attempt diverged early despite identical initial seeding, failing to reproduce the baseline pattern. Neither conversational length nor memory presence reliably predicted outcomes.

5. Discussion

This study set out to document a narrow set of observable conversational behaviors across a small number of extended interactions with large language models. The analysis was explicitly descriptive in scope and avoided claims about internal model states, mechanisms, or subjective experience. The results indicate that the classified behaviors were conditional and interaction-dependent rather than generic properties of extended dialogue. However, this work does not identify which specific interaction features, if any, are necessary or sufficient for their appearance. The findings therefore constrain the generality of the baseline observation without establishing explanatory mechanisms. The failure of partial replication seeded with identical initial prompts demonstrates that similar starting conditions are insufficient to reliably reproduce the observed pattern, reinforcing the probabilistic and context-sensitive nature of the phenomenon. Throughout the analysis, care was taken to distinguish between surface-level conversational features and the more structurally distinctive patterns captured by the classification criteria. Language indicating continuation or engagement was treated as a behavioral feature of the interaction rather than evidence of intent, preference, or internal motivation.

6. Limitations and Future Work

This study is limited by its small sample size and reliance on interactions conducted by a single user. Because the same individual conducted the interactions and performed the classification, subtle prompting biases or interpretive expectations cannot be ruled out. Future work could employ larger

corpora, independent annotation, and more controlled experimental designs.

Appendix A: Scope and Non-Claims

This paper makes no claims regarding consciousness, subjective experience, intentionality, agency, or internal cognitive states of large language models. All findings are descriptive and pertain solely to observable interaction dynamics.

Author's Note on Methodology and Use of AI Tools

AI-assisted tools were used for drafting and editing. These tools were not used to generate data, select evidence, define criteria, or determine conclusions. All analytical decisions reflect the author's intent and oversight.

References

Anthropic. (2023). *Anthropic research overview.*
https://www.anthropic.com/research

Anthropic. (2022). *Constitutional AI: Harmlessness from AI feedback.*
https://www.anthropic.com/research/constitutional-ai

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big?* Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.
https://doi.org/10.1145/3442188.3445922

Weidinger, L., et al. (2021). *Ethical and social risks of harm from language models.* arXiv preprint arXiv:2112.04359.
https://arxiv.org/abs/2112.04359

Bommasani, R., et al. (2021). *On the opportunities and risks of foundation models.* arXiv preprint arXiv:2108.07258.
https://arxiv.org/abs/2108.07258