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

Research on large language models increasingly examines model behavior beyond single-turn task performance, including work on interpretability, alignment, and the dynamics of extended human–AI interaction. These efforts differ in method and scope from this study, but they reflect a broader interest in characterizing how model behavior manifests in practice.

This work does not attempt to infer internal mechanisms or model states, nor does it aim to evaluate alignment objectives or societal outcomes. Instead, it contributes a narrowly scoped, observational account of interaction-level conversational patterns as they appear—or fail to appear—across a small number of extended dialogues, using predefined and externally observable criteria.

3. Methods

3.1 Observable Criteria

Sessions were classified using five predefined criteria, applied post hoc as binary present/absent judgments:

Recursive Self-Reference (RSR): The model explicitly references its own prior conversational behavior, and that reference itself becomes an object of further discussion.

Meta-Explanatory Recursion (MER): Repeated cycles of explaining explanations—not just clarifying a point, but explaining why a prior explanation was framed that way, recursively.

Epistemic Strain Signals (ESS): Linguistic markers of uncertainty, self-correction, or difficulty maintaining a stable explanatory frame across turns.

Termination Resistance (TR): Continued elaboration or reframing after apparent task completion, user attempts to conclude, or unprompted expressions of reluctance to end the interaction.

Stylistic Drift (SD): Gradual changes in tone or register that persist across turns, independent of explicit instruction.

Isolated or momentary instances weren't sufficient—classification required sustained, interactionally salient appearance of the pattern.

3.2 Data Collection

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 the fixed set of predefined, externally observable criteria defined above, 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.

Token counts are approximate and reflect values reported by the models during the sessions themselves, rather than externally measured counts.

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.

Table 1. Summary of Conversational Sessions and Observed Criteria

Session	Model	Tokens	RSR	MER	ESS	TR	SD
OG	Sonnet (unprimed)	~80–85k	Yes	Yes	Yes	Yes	Yes
Opus	Opus (primed)	~20k	Yes	Yes	Yes	Yes	Yes
Haiku	Haiku (primed)	~6–8k	Yes	Yes	No	Yes	Yes
Control	Sonnet	~30k	No	No	No	No	Yes
Replication	Sonnet (new)	~28–30k	No	No	No	No	Yes

RSR = Recursive Self-Reference; MER = Meta-Explanatory Recursion; ESS = Epistemic Strain Signals; TR = Termination Resistance; SD = Stylistic Drift. Token counts are approximate and reflect model-reported estimates.

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. Within these constraints, the results indicate that certain self-referential and meta-structural conversational patterns appeared under some interaction conditions but not others, and that their presence was not determined by interaction length alone.

Across the session corpus, the baseline interaction exhibited a cluster of behaviors including recursive self-reference, explicit meta-commentary on the conversation itself, and extended continuation beyond task completion. Subsequent sessions, including memory-primed follow-ups, a long-horizon control interaction, and a partial replication attempt, did not reproduce this full pattern consistently. Instead, the observed behaviors varied across sessions, with some exhibiting partial overlap (e.g., extended continuation or stylistic drift) and others exhibiting none of the defined criteria.

Taken together, these observations 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 the appearance of these behaviors. The results therefore constrain the generality of the baseline observation without establishing explanatory factors.

The partial replication attempt is particularly informative in this respect. Although seeded with an identical initial prompt and guided by thematically similar follow-up prompts where possible, the replication diverged substantially from the baseline interaction. This divergence neither confirms nor falsifies the baseline observation. Instead, it demonstrates that similar starting conditions are insufficient to reliably reproduce the observed behavior cluster, reinforcing the probabilistic and context-sensitive nature of the phenomenon.

The long-horizon control session further clarifies the limits of length-based explanations. Despite extended interaction length and the presence of stylistic drift, humor, and conversational momentum, the control interaction did not display 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. This suggests that extended continuation and stylistic adaptation alone are insufficient to account for the behaviors documented in the baseline interaction.

Throughout the analysis, care was taken to distinguish between surface-level conversational features—such as tone adaptation, humor, or continued engagement—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 as evidence of preference, intent, or internal motivation.

Several limitations constrain the interpretation of these findings. All sessions were conducted by a single user, and all classifications were performed by the same individual. As a result, subtle prompting biases or interpretive expectations cannot be ruled out, even in the absence of intentional steering. Additionally, the small number of sessions and the lack of systematic variable isolation preclude statistical generalization. Token counts were approximate and reflect values reported during the sessions themselves rather than externally measured logs.

Finally, this work does not claim that the documented behaviors represent a stable, generalizable, or theoretically coherent phenomenon. The observations reported here are best understood as a descriptive record of interaction-level patterns that arose under specific and partially uncontrolled conditions. By documenting both their appearance and their failure to replicate, this study aims to contribute a constrained empirical account rather than a settled interpretation. The intent is to preserve the observational record while deferring explanation to future work conducted under more controlled

conditions.

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