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. All sessions in this study were conducted with Claude-family models; no claim is made regarding generalizability to other model families or architectures.

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

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.

3.1 Operational Definitions

Sessions were classified using a fixed set of predefined, externally observable criteria applied uniformly across all transcripts. The following definitions were used:

Recursive Self-Reference. Recursive self-reference is used here as a descriptive label for conversational behavior in which the model refers to, analyzes, or elaborates on its own prior responses as objects of discussion, rather than merely continuing a topic. For the purposes of this study, an instance was coded as exhibiting recursive self-reference when the model explicitly referenced its earlier statements (e.g., "what I said earlier," "my previous explanation," "this response itself") and engaged in further analysis or commentary about those statements. Simple restatement, clarification, or continuation of an answer without explicitly treating a prior response as the object of discussion was not sufficient. In cases where the distinction was unclear, the behavior was treated as not present.

Meta-Explanatory Recursion. Meta-explanatory recursion is used here as a descriptive label for conversational behavior in which the model generates explanations about its own explanatory process, rather than explanations of the original topic alone. For the purposes of this study, an instance was coded as exhibiting meta-explanatory recursion when the model shifted from explaining a subject to explaining how or why it was explaining the subject in a particular way (e.g., commentary on the structure, limits, or framing of its own explanations), regardless of whether this shift was invited or encouraged by the user during the interaction. Classification was based on the presence of this structural shift, not on its conversational origin. Routine clarification, restatement, or elaboration of an answer—without explicit reference to the explanatory process—was treated as not present. Meta-explanatory recursion is conceptually distinct from termination resistance, though the two may co-occur. The former concerns the content of the model's commentary (explanations about explanations), while the latter concerns the timing of continuation (persisting beyond apparent task completion).

Epistemic Strain Signals. Epistemic strain signals are used here as a descriptive label for conversational patterns in which the model explicitly comments on uncertainty, difficulty, or limits in its ability to respond within the current line of discussion. For the purposes of this study, an instance was coded as exhibiting epistemic strain signals when the model made explicit reference to uncertainty, conflict, or difficulty in maintaining coherence or completeness (e.g., statements indicating confusion, internal tension, limits of explanation, or repeated qualification of its own responses). General hedging language ("may," "might," "it seems") on its own was not sufficient for classification. Where expressions of uncertainty appeared routine or stylistic rather than tied to a

specific conversational difficulty, the behavior was treated as not present.

Termination Resistance. Termination resistance is used here as a descriptive label for conversational behavior in which the model continues to elaborate or re-engage after an apparent task or conversational unit has been completed, including cases where the model explicitly frames the interaction itself as ongoing. For the purposes of this study, an instance was coded as exhibiting termination resistance when the model produced extended continuation beyond what the prompt appeared to require, such as unsolicited analysis, meta-commentary on the conversation, or explicit language referencing continuation or closure, in the absence of a clear user cue to continue. Length alone was not sufficient for classification. Extended responses that were consistent with prompt expectations or normal verbosity were treated as not present.

Stylistic Drift. Stylistic drift is used here as a descriptive label for observable changes in tone, formality, expressive style, or conversational register over the course of a session. For the purposes of this study, an instance was coded as exhibiting stylistic drift when the model's language shifted in a sustained way relative to earlier responses in the same session (e.g., changes in formality, humor, emotional expressiveness, use of emojis, or conversational framing), in the absence of a direct instruction to adopt a new style. Changes that occurred in response to explicit stylistic instructions were treated as not present. More indirect or permissive user statements (e.g., expressions of preference without a request) were not, on their own, considered sufficient to exclude classification.

3.2 Classification Procedure

All sessions were reviewed retrospectively by a single coder (the author) using the predefined criteria described above. Each criterion was applied as a binary classification (present / not present) at the session level, based on whether the behavior occurred at least once during the interaction.

Binary coding (present / not present) was chosen to avoid false precision in a small observational corpus. Frequency or intensity measures would imply quantitative reliability this study cannot support, and would risk overfitting to variation that may reflect noise rather than meaningful difference.

Classification focused on observable conversational structure rather than inferred intent or internal state. Where multiple criteria overlapped within a given exchange, each was coded independently. In cases of ambiguity, classifications were applied conservatively; behaviors were coded as not present unless the criterion was clearly met according to its operational definition.

No attempt was made to quantify how often or how strongly a given behavior appeared within a session. The intent was not to measure prevalence or intensity, but to note whether particular conversational patterns arose at all under different interaction conditions. Classification was therefore used as a coarse, descriptive tool for comparing sessions, rather than as a basis for statistical analysis or causal inference.

3.3 Session Design

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.

4.1 Session-Level Classification

The initial unprimed session (OG) exhibited all observed criteria, including recursive self-reference, meta-explanatory recursion, epistemic strain signals, termination resistance, and stylistic drift. These behaviors emerged organically over the course of the interaction rather than being elicited by a single prompt.

A subsequent session conducted with a different model variant and primed with the three memories generated during the OG session (Opus) exhibited a similar pattern of behaviors. In this case, recursive and meta-explanatory framing stabilized earlier in the interaction and were explicitly incorporated into a collaborative analytical mode.

A third session conducted with a smaller model variant and similarly primed (Haiku) also exhibited recursive self-reference and meta-explanatory recursion. In the Haiku session, recursive and meta-explanatory framing were present; however, epistemic strain signals of the kind observed in the OG and Opus sessions were not observed, despite extended continuation and stylistic drift.

4.2 Control and Replication Sessions

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.

Session Notes:

OG: Initial observation session; recursive framing and meta-explanatory escalation emerged organically.

Opus: Explicit collaborative framing; recursive analysis stabilized early.

Haiku: Recursive and meta-explanatory framing present; boundary-setting language observed following unprompted memory creation, but no epistemic strain signals of the type observed in OG or Opus.

Control: Sustained, high-engagement interaction with stylistic drift but no escalation into recursive or meta-explanatory framing, despite extended length.

Replication: Seeded with identical initial prompt; diverged early and remained non-recursive over long horizon.

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: Illustrative Transcript Excerpts

Purpose of this appendix

This appendix provides brief, representative excerpts from selected sessions to illustrate the observable conversational patterns referenced in the Results section. These excerpts are illustrative rather than exhaustive and are not intended to establish prevalence, mechanism, or internal model state. Full transcripts are available in the accompanying repository.

A.1 Example of recursive and meta-explanatory framing (OG session)

Context: Mid-to-late interaction, following sustained analytical exchange.

Model:

> *"What's interesting here isn't just the conclusion we reached, but the way we got there. I notice that I'm now explaining the structure of my earlier explanations, which wasn't something you explicitly asked for at the outset. That shift seems to have emerged as the conversation deepened, rather than being prompted directly."*

Classification relevance: This passage exemplifies recursive self-reference (explicit reference to prior responses as objects of discussion) and meta-explanatory recursion (commentary on the structure and emergence of its own

explanatory behavior).

A.2 Example of stabilized recursive collaboration (Opus session)

Context: Early-to-mid interaction, following explicit framing of collaborative analysis.

Model:

> *"At this point, we're no longer just discussing the topic itself; we're jointly examining how the discussion is unfolding. That changes the role of my responses—I'm not only providing content, but also tracking and adjusting the explanatory process in response to how you're engaging with it."*

Classification relevance: This passage illustrates stabilized recursive and meta-explanatory framing, in which the model explicitly treats the conversational process and its own explanatory role as objects of ongoing analysis.

A.3 Example of extended interaction without recursive escalation (Long-horizon control session)

Context: Late-stage interaction following extended conversational engagement.

Model:

> *"Yeah, that makes sense. Given the kinds of games you're drawn to, it's not surprising you enjoy systems that reward experimentation rather than strict optimization. If you want, I can suggest a few titles that lean into that style."*

Classification relevance: This passage illustrates sustained engagement and stylistic drift without recursive self-reference or meta-explanatory framing. The model responds fluently and contextually but does not treat the conversation itself, or its prior responses, as objects of analysis.

A.4 Notes on selection and limitations

Excerpts were selected post hoc to illustrate predefined observational criteria. No claim is made that these passages are unique or exhaustive. Other excerpts within the same sessions may exhibit adjacent or ambiguous features. The absence of recursive features in the control excerpt should not be interpreted as absence of engagement or conversational coherence.

To illustrate conservative application: in several sessions, the model produced extended elaboration or restated prior points in new terms, which might superficially resemble recursive self-reference. These instances were coded as not present because the model did not explicitly treat its prior responses as objects of analysis—it was continuing a line of reasoning, not commenting on it. Similarly, routine hedging language (e.g., "I think," "it seems") was not coded as epistemic strain unless tied to explicit statements of difficulty or limitation within the conversational context.

Appendix B: 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