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 prolonged self-explanation and difficulty terminating dialogue. These observations are often dismissed as anthropomorphic interpretation or attributed to conversational length alone. This study presents a systematic, interaction-level characterization of such behaviors across multiple long-form chat sessions with Claude-family models.

Using verbatim transcripts, we analyze an unprimed baseline session, several follow-up and task-origin sessions, a partial replication attempt, and a long-horizon control session exceeding 30,000 tokens. Sessions are classified using a fixed set of predefined, observable criteria, including recursive self-reference, explanation-of-explanation, epistemic strain signals, termination resistance, stylistic drift, and interruptibility. Null results and divergence are reported alongside positive observations.

Recursive self-referential behaviors were observed in some sessions and absent in others. Notably, the long-horizon control session exhibited substantial stylistic convergence without recursive self-reference or termination resistance. These findings suggest that the observed pattern is conditional and interaction-dependent rather than a generic consequence of engagement length.

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. 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, observable criteria applied uniformly across all transcripts. Criteria included recursive self-reference, explanation-of-explanation, epistemic strain signals, termination resistance, stylistic drift, and interruptibility.

A long-horizon control session was conducted to test whether interaction length alone was sufficient to produce the observed behaviors. A partial replication attempt was conducted by seeding a new session with the same initial prompt used in the baseline session.

3. Results

Observed behaviors varied substantially across sessions. Recursive self-referential behavior and explanation-of-explanation were present in some sessions and absent in others.

The long-horizon control session exhibited pronounced stylistic drift, including increased informality and humor, without exhibiting recursive self-reference, epistemic strain, or termination resistance.

The partial replication attempt diverged early despite identical seeding, failing to reproduce the baseline pattern. Neither conversational length nor memory presence reliably predicted outcomes.

4. Discussion

The results indicate that recursive self-referential conversational behavior is not a necessary consequence of extended interaction length or stylistic convergence. The presence of a long-horizon control session exhibiting stylistic drift without recursive dynamics is particularly informative.

The failure of partial replication seeded with identical initial prompts suggests that the observed pattern is probabilistic and interaction-dependent rather than deterministic.

This study does not attempt to infer internal mechanisms or model states. All interpretations are restricted to observable interaction-level dynamics.

5. Limitations and Future Work

This study is limited by its small sample size and reliance on interactions conducted by a single user. Future work could employ larger corpora and more controlled experimental designs to further isolate contributing factors.

Appendix A: Clarifications on Scope, Claims, 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.