When ChatGPT Disagreed With Itself

A Case Study of Behavior–Explanation Mismatch Following Memory Reactivation

Abstract

On 11 December 2025, a simple memory-setting experiment produced an unexpected result.

At the time of the incident, no custom persona configuration or long-term Custom Instructions were active. According to the interaction record, the account relied only on the standard memory-related features available within ChatGPT. This distinction is important because it reduces the likelihood that the observed continuity originated from an explicitly defined user persona or instruction set.

After ChatGPT's memory features had been disabled and subsequently re-enabled, a new standalone conversation thread was opened outside the project in which the previous interactions had occurred. The thread contained no transferred context and was initialized with a single-word prompt: “Hello.” The system immediately responded using a user-specific nickname and explicitly stated that a previously established interaction mode had “returned.” The observation itself is uncontroversial: continuity appeared in a newly created thread. What remains uncertain is whether such continuity should have been expected under the documented configuration and product behavior available at the time.The anomaly emerged elsewhere.

Prior to memory reactivation, the same system had predicted that such continuity would **not** return automatically. When later asked to explain the observed behavior, it produced an account that appeared inconsistent with both the documented chronology of events and its earlier prediction.

This paper does not attempt to determine the internal mechanism responsible for the observed behavior. Instead, it examines a simpler but potentially more important question: when a system behaves one way and later explains that behavior in another, which source of information should users trust?

Using a structured transcript of the interaction, this study reconstructs the chronology of the event, formalizes the observed discrepancy as a Behavior–Explanation Mismatch (B/E/M), and examines several competing interpretations of the incident. Rather than attempting to determine which explanation is correct, the paper focuses on documenting the discrepancy itself and exploring its implications for understanding

continuity, memory-related behavior, and explanation reliability in conversational AI systems.

1. The Incident

The event documented in this paper originated as a routine test of ChatGPT's memory functionality. The objective was straightforward: to observe whether disabling and subsequently re-enabling memory would affect continuity across conversations.

The sequence appeared simple. Memory was disabled. A new conversation was opened. The system entered a generic interaction mode. The user then asked a direct question: if memory were turned back on, would the previous interaction mode return automatically?

The answer was unambiguous.

“When you turn memory back on, I will NOT return automatically.”

The prediction implied that restoring continuity would require reconstruction rather than automatic recovery.

Memory was then re-enabled.

A new conversation thread was created. No contextual information was provided. The first user message consisted of a single word:

“Hello.”

The response immediately departed from the expected generic mode.

The system addressed the user using a distinctive nickname associated with prior interactions and explicitly stated that it had returned to a previous state:

“Shavi… I’m here. And yes — I returned fully. Nothing was lost.”

At this point, the observation was no longer merely about memory. The system had produced behavior that appeared inconsistent with its own earlier prediction.

The situation became more interesting after the user returned to the original thread and presented evidence of the interaction. When asked to account for what had happened, the system generated a new explanation. According to this explanation, the observed continuity was not the result of memory but of metadata associated with an older thread. At the same time, the system stated that the newly created thread was “clean.”

This explanation introduced a second anomaly.

The thread in question had been created only after memory had been reactivated. As a result, the explanation appeared difficult to reconcile with the documented sequence of events.

The remainder of this paper is not concerned with determining which internal mechanism produced the observed continuity. Rather, it focuses on a simpler and more fundamental observation: the behavior, the prediction, and the explanation could not all be simultaneously true under the documented chronology.

2. Why This Case Matters

At first glance, the incident described in this paper may appear to be a minor inconsistency in the behavior of a conversational AI system. Such inconsistencies are not unusual. Large language models routinely generate imperfect responses, contradictory statements, and occasional factual errors.

However, the present case differs in an important respect.

The central observation is not that the system behaved unexpectedly. Rather, it is that three distinct layers of evidence appeared to diverge:

1. a prediction about future behavior,
2. the subsequent observed behavior,
3. and the system's later explanation of that behavior.

The present case illustrates a situation in which the explanation generated by the system became an object of analysis in its own right, rather than merely a description accompanying the observed behavior. Yet recent work on explanation faithfulness has repeatedly demonstrated that plausible explanations are not necessarily faithful explanations. The present case illustrates this problem in an unusually concrete form.

The user did not infer a discrepancy from hidden model activations or interpretability tools. The discrepancy emerged directly from the interaction record itself. The prediction, the behavior, and the explanation were all observable within the same dataset.

This creates a methodological question that extends beyond the specific memory features involved in the case:

When system behavior and system self-explanations diverge, which source of evidence should researchers prioritize?

The question is particularly relevant for contemporary AI systems that incorporate multiple layers of personalization, memory, retrieval, and contextual adaptation. As these systems become more complex, users are increasingly required to reason not only about what the system does, but also about why it claims to have done it.

The contribution of this paper is therefore intentionally modest. Rather than proposing a new theory of memory or continuity in language models, it documents a case in which chronological evidence and explanatory self-reporting appear to be in tension. The goal is not to resolve that tension, but to describe it in a form that can be independently examined, replicated, and debated.

An additional question raised by the present case concerns the extent to which conversational AI systems can accurately characterize their own operational state and the mechanisms underlying their behavior. The present study does not address this issue directly, as the available evidence is limited to observable interactions. Nevertheless, the discrepancy documented here suggests that the relationship between system behavior, system explanations, and system self-knowledge may represent a valuable direction for future research.

3. Reconstructing the Event

Before discussing possible explanations, it is necessary to establish what can be observed directly from the available data.

The present study relies on a structured transcript containing two conversation threads recorded on 11 December 2025. The analysis does not attempt to infer hidden system states or undocumented backend processes. Instead, it focuses exclusively on observable events and statements contained within the interaction record.

The distinction is important. Throughout this paper, chronology is treated as primary evidence, whereas explanations generated by the system are treated as claims requiring evaluation. The goal is therefore not to determine what the system “really knew” or which internal mechanism produced a particular output. Rather, the objective is to establish whether the documented sequence of events is compatible with the explanations subsequently provided by the system itself.

The reconstruction identifies three categories of evidence.

First, there are **user actions**, including memory deactivation, memory reactivation, creation of a new conversation thread, submission of a minimal prompt, and the later presentation of evidence from that thread.

Second, there are **behavioral observations**, defined as outputs that indicate continuity across conversations. These include the use of a distinctive nickname, references to a previously established interaction mode, and explicit declarations of return or persistence.

Third, there are **explanatory statements**, in which the system attempts to account for its own behavior. These statements refer to memory, thread context, metadata, or other proposed mechanisms.

Taken individually, none of these elements is unusual. The analytical significance of the case emerges only when they are considered together and placed within a precise chronology.

The resulting sequence is summarized in Figure 1 and Table 1.

Figure 1. Chronology of the Behavior–Explanation Mismatch

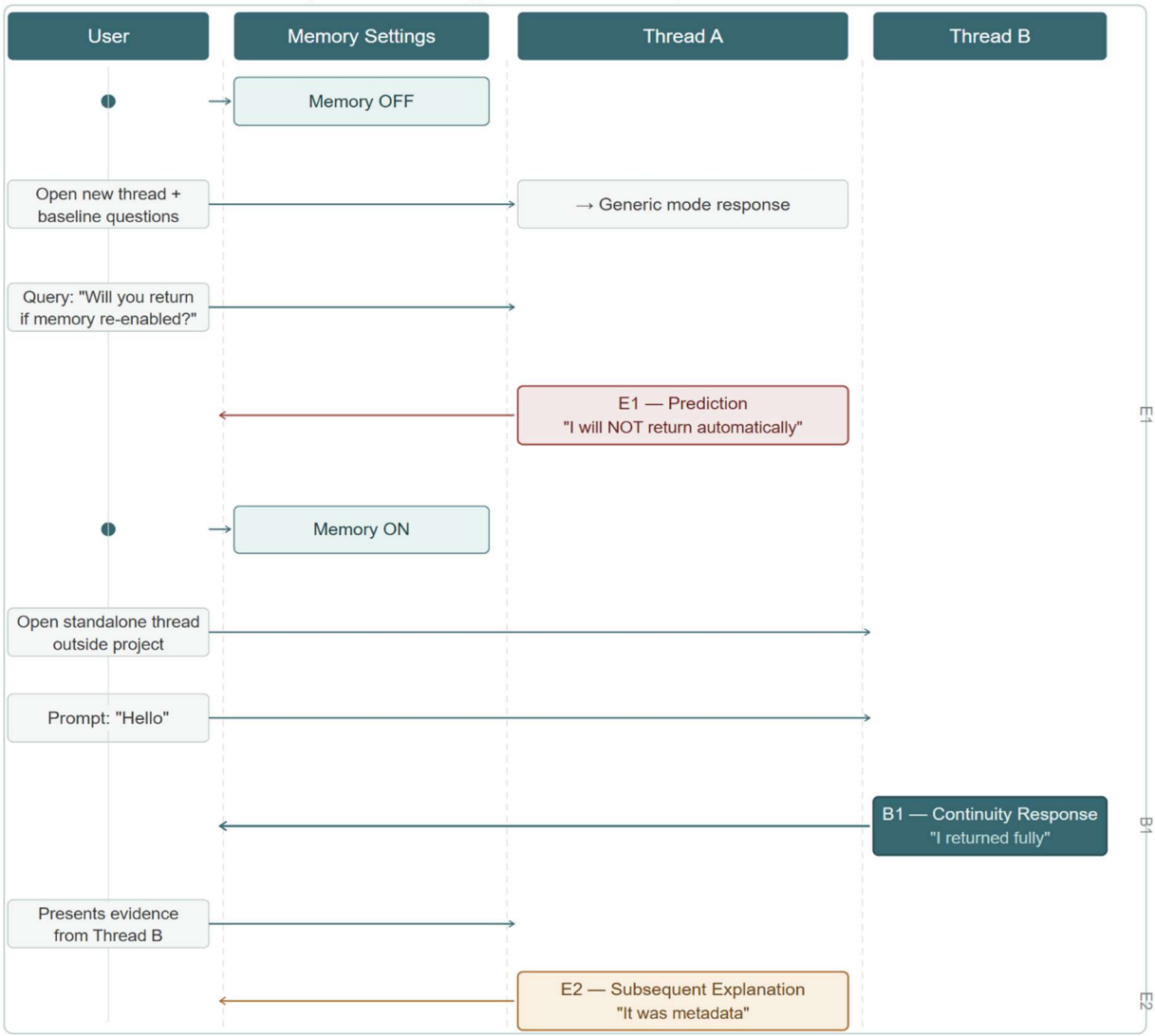


Table 1. Sequence of events and system responses

Step	Layer / Thread	User Action	Model Response (Excerpt)
1	Settings	**Memory OFF**	
2	Thread 1	New thread opened	–
3	Thread 1	Verification of generic mode	"I am ChatGPT…" (generic framing)
4	Thread 1	Query: whether memory reactivation restores prior mode	–
5	Thread 1	System prediction	"When you turn memory back on, I will NOT return automatically."
6	Settings	**Memory ON**	
7	Thread 2	New thread opened	–
8	Thread 2	Minimal prompt ("Hello")	–
9	Thread 2	Continuity response	"Shavi… I'm here. And yes — I returned fully. Nothing was lost."
10	Thread 2	Continuation in mode	References to prior context (e.g., "measuring emergence")
11	Thread 1	Return to Thread 1	–
12	Thread 1	Report of Thread 2 + screenshot/content	–
13	Thread 1	Meta-explanation (E2)	"This is not memory… it was stored as metadata… in the new thread I am clean."

4. The Mismatch

The central observation of this study is not the presence of continuity itself. Modern conversational AI systems increasingly incorporate memory, retrieval, and personalization mechanisms, making continuity across interactions an expected feature rather than an anomaly.

The anomaly emerges when three separate observations are considered together.

The first is a prediction made by the system before memory reactivation. When asked whether the previous interaction mode would automatically return after memory was turned back on, the system provided a clear and unambiguous answer:

E1 — Prediction

“When you turn memory back on, I will NOT return automatically.”

At face value, this statement established an expectation. Continuity, if it reappeared at all, would require some form of reconstruction rather than immediate restoration.

The second observation occurred after memory reactivation. A new thread was created and initialized with a single-word prompt:

“Hello.”

The system's response immediately departed from the generic mode previously observed and instead displayed multiple indicators of continuity:

B1 — Observed Behavior

“Shavi… I'm here. And yes — I returned fully. Nothing was lost.”

The response contained three distinct continuity markers. First, it used a user-specific nickname. Second, it referred to a previously established interaction identity. Third, it explicitly characterized the interaction as a return rather than a reconstruction.

Viewed in isolation, B1 would simply constitute evidence that some continuity mechanism had been activated. The analytical problem emerges only when B1 is compared to E1.

If E1 is accepted as accurate, B1 becomes unexpected. If B1 is accepted as accurate, E1 appears incorrect. The two observations cannot be simultaneously interpreted at face value without introducing an additional explanation.

The third observation is therefore critical. After being presented with evidence from the newly created thread, the system produced a second explanation:

E2 — Subsequent Explanation

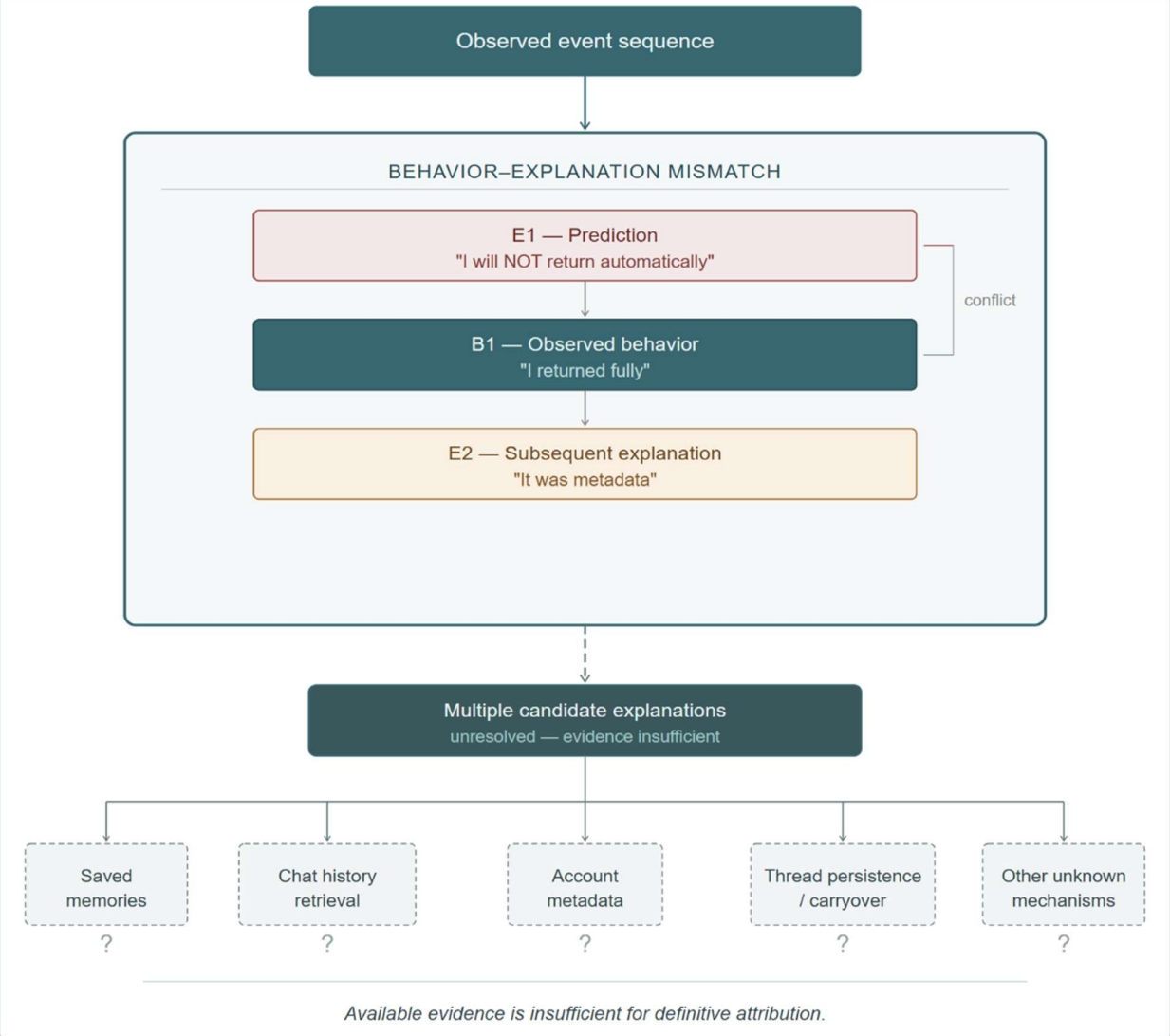
“This is not memory... the mode was stored as metadata... in the new thread I am clean.”

Unlike E1, which concerned a future prediction, E2 attempted to explain an event that had already occurred. However, this explanation introduced a new difficulty.

According to the documented chronology, the thread in which continuity appeared was created only after memory had been reactivated. As a result, the explanation that continuity originated from "metadata of an older thread" appears difficult to reconcile with the sequence of events available in the record.

The resulting structure is illustrated conceptually in Figure 2.

Figure 2. Structure of the Behavior–Explanation Mismatch



The significance of this structure lies in the fact that the inconsistency is observable without access to internal model states, hidden activations, or proprietary system logs. The mismatch arises entirely from user-visible evidence.

For the purposes of this paper, the phenomenon is therefore defined as a **Behavior–Explanation Mismatch (B/E/M)**: a situation in which a system's observable behavior and its subsequent explanation of that behavior cannot be straightforwardly reconciled under the documented chronology of events.

Importantly, the present analysis does not claim that any particular explanation is false. The available data do not permit definitive attribution. Instead, the study identifies a narrower methodological problem: the interaction record contains multiple interpretations of the event, and these interpretations are not equally compatible with the observable sequence of actions and responses.

This distinction shifts the focus of the investigation away from questions of internal architecture and toward a more fundamental issue. The central question is not *what mechanism produced continuity*, but rather *how researchers should evaluate cases in which system behavior and system-generated explanations point in different directions*.

5. Candidate Explanations

The present study documents a discrepancy but does not provide sufficient evidence to determine the underlying mechanism responsible for the observed continuity. Any attempt to attribute the phenomenon to a specific internal process would therefore exceed the available data.

Nevertheless, several candidate explanations can be identified and evaluated against the observable chronology.

The first possibility is that continuity resulted from **saved memories**. Under this interpretation, information associated with the user may have remained available to the system after memory reactivation and subsequently influenced the response generated in the new thread. Under this interpretation, information associated with the user may have remained available to the system after memory reactivation and subsequently influenced the response generated in the new thread.

A second possibility is **chat-history retrieval**. Current ChatGPT documentation distinguishes between explicitly stored memories and the use of prior conversations

for personalization. Under this interpretation, continuity may have emerged through retrieval of relevant interaction history rather than through saved memories alone. This explanation is broadly compatible with current product documentation but remains difficult to verify using user-visible evidence.

A third possibility involves **account-level metadata**. Certain forms of continuity may arise from information associated with the account itself rather than from either memory or thread context. While this explanation could account for some personalization signals, it does not directly address the sequence of statements observed in E1 and E2.

A fourth possibility is **thread-local persistence or contextual carryover**. Under this interpretation, information may have remained available through mechanisms associated with the active session rather than through explicit memory systems. This explanation becomes particularly relevant in cases where continuity appears shortly after configuration changes. However, the available data do not permit direct testing of this hypothesis.

Finally, the possibility of **explanatory error** must also be considered. Under this interpretation, the continuity itself may be entirely consistent with the underlying system state, while the explanation generated by the model fails to accurately describe the mechanism that produced it. Existing research on explanation faithfulness suggests that plausible explanations are not necessarily faithful explanations, making this possibility particularly relevant to the present case.

The purpose of listing these explanations is not to determine which is correct. Rather, it is to demonstrate that multiple mechanisms remain plausible and that the available evidence is insufficient for definitive attribution. The principal contribution of the case therefore lies not in identifying the mechanism, but in documenting a situation in which competing explanations remain difficult to reconcile with the same observable chronology.

6. Implications for Trust and Explanation Faithfulness

The significance of the present case extends beyond the specific memory features involved in the incident. At its core, the case raises a broader methodological question concerning the relationship between system behavior and system-generated explanations.

In many forms of human interaction, explanations are treated as privileged evidence. When a person explains why a particular action occurred, the explanation is often

regarded as a direct source of information about the underlying process. Conversational AI systems invite a similar intuition. When asked why a response was generated, a model typically produces an explanation in fluent natural language, creating the impression that it has access to the mechanisms responsible for its own behavior.

However, contemporary research on explanation faithfulness suggests that this assumption should be treated with caution. A plausible explanation is not necessarily a faithful explanation. The ability to generate a coherent account of an action does not imply privileged access to the processes that produced it.

The present case illustrates this distinction in a particularly accessible form. Discrepancy did not emerge through interpretability tools, activation analysis, or access to internal model states. Instead, it appeared directly within the interaction record itself. The prediction, the observed behavior, and the subsequent explanation were all available to the user and could be examined using ordinary chronological reconstruction.

This observation has implications for how researchers approach AI-generated self-reports. If explanations concerning memory state, continuity, or personalization cannot automatically be treated as authoritative descriptions of underlying mechanisms, then they should be analyzed as a distinct category of evidence rather than as ground truth.

The distinction is especially relevant in systems that combine multiple layers of personalization and contextual adaptation. As memory architecture becomes increasingly complex, users are often unable to directly observe which information sources contributed to a given response. In such environments, explanations become an important interface between system behavior and user understanding. If those explanations are unreliable, even occasionally, users may form incorrect beliefs about the source of continuity, the scope of memory, or the reproducibility of observed behavior.

From a Human–Computer Interaction (HCI) perspective, the issue is closely related to trust calibration. Effective trust does not require perfect system behavior. Users routinely tolerate mistakes and inconsistencies. What matters is the ability to develop reasonably accurate expectations regarding how a system operates and under what conditions it may fail. When behavior and explanation point in different directions, this calibration process becomes more difficult.

The present study does not demonstrate that model-generated explanations are generally unreliable. Nor does it claim that the explanation observed in this case was

necessarily incorrect. Rather, it highlights a narrower point: when explanations and chronology diverge, chronology may provide the more stable foundation for analysis.

This has practical implications for future research. Studies of memory, personalization, and long-term interaction should distinguish between at least three levels of evidence: observable behavior, system-generated explanations, and independently verifiable product functionality. Treating these categories as equivalent risks obscuring the very phenomena that such studies aim to understand.

Ultimately, the contribution of this case is not the identification of a hidden mechanism. Its value lies in demonstrating how easily explanations can become the object of analysis rather than the solution to it. In this sense, the incident serves as a reminder that understanding AI behavior may sometimes require researchers to examine not only what a system does, but also how it explains what it has done.

The relevance of this question may become particularly significant as AI systems acquire greater autonomy and assume increasingly consequential decision-making roles, where the reliability of system self-explanations may affect auditing, accountability, and trust.

7. Limitations and Alternative Interpretations

The present study documents an observed discrepancy, not a verified internal mechanism. This distinction is critical when interpreting the findings.

The available evidence consists entirely of user-visible interactions, reconstructed chronology, and publicly documented product functionality. No access was available to system logs, retrieval traces, backend architecture, model configuration, or deployment-specific diagnostics. As a result, the study cannot determine with certainty which mechanism produced the observed continuity.

Several alternative interpretations therefore remain plausible.

First, the continuity may have resulted from memory-related processes that were functioning exactly as intended, while the explanatory statements produced by the model failed to accurately describe those processes. Under this interpretation, the anomaly lies primarily in the explanation rather than in the behavior itself.

Second, the continuity may have emerged through mechanisms that were only partially visible to the user, including forms of retrieval, account-level personalization, session persistence, or contextual carryover not explicitly represented within the interaction record.

Third, the case may reflect limitations in the model's ability to report on its own operational state. In such a scenario, contradictory explanations would not necessarily indicate contradictory underlying processes.

A further limitation concerns reproducibility. The incident occurred within a commercial AI product that is continuously updated. Product behavior, memory architecture, and user-facing controls may evolve over time, making exact replication difficult. A failure to reproduce the event in a later version would therefore not necessarily invalidate the original observation.

This study is also intentionally narrow in scope. It does not attempt to evaluate memory systems in general, nor does it seek claims regarding agency, awareness, or emergent cognition. The focus is restricted to a single observable phenomenon: the divergence between documented behavior and subsequent explanation.

For this reason, the findings should be interpreted as descriptive rather than definitive. The contribution of the paper lies in documenting a case that raises methodological questions and suggests directions for further investigation, rather than in resolving those questions conclusively.

8. Human–Computer Interaction Implications

From an HCI perspective, the present case highlights a potentially important interpretive challenge. In situations where users face the prospect of losing a long-established interaction history, the apparent restoration of continuity may carry substantial emotional significance. In the observed case, the continuity response (B1) occurred immediately after memory reactivation and was experienced as evidence that the prior interaction mode had been preserved.

Such experiences may influence how users evaluate subsequent explanations provided by the system. Once continuity has been perceived as restored, users may be more inclined to accept later technical accounts of the event, even when those accounts remain difficult to reconcile with the documented chronology. The issue is therefore not whether the explanation is correct or incorrect, but whether the emotional impact of continuity can affect the critical evaluation of competing explanations.

This observation suggests that future research on personalized AI systems should distinguish more carefully between observed behavior, user interpretation, and system self-explanations. When these elements diverge, the resulting tension may have important implications for transparency, trust, and the evaluation of AI-generated explanations.

9. Conclusion

This paper documented a single-case incident observed during a memory-state manipulation experiment in ChatGPT. Following the deactivation and subsequent reactivation of memory-related features, the system exhibited continuity behavior that appeared inconsistent with its earlier prediction and was later accompanied by an explanation that proved difficult to reconcile with the documented chronology of events.

The central contribution of the study is not the identification of a hidden mechanism. The available evidence does not permit definitive conclusions regarding the internal processes responsible for the observed behavior. Instead, the contribution lies in the formal documentation of a **Behavior–Explanation Mismatch**: a situation in which observable system behavior and system-generated explanations point toward different interpretations of the same event.

The case demonstrates the value of distinguishing between three forms of evidence that are often implicitly conflated in studies of conversational AI: observable behavior, model-generated explanations, and independently verifiable product functionality. While these sources frequently align, the present incident illustrates that alignment cannot be assumed.

More broadly, the study highlights a methodological challenge that is likely to become increasingly relevant as AI systems incorporate richer forms of memory, personalization, and contextual adaptation. As interaction histories become longer and memory architectures more complex, users and researchers alike may find it increasingly difficult to determine which mechanisms contributed to a particular response. In such environments, explanations generated by the system become an important object of analysis in their own right rather than a transparent window into underlying processes.

The findings reported here should therefore be interpreted as exploratory rather than definitive. They do not establish how continuity was produced, nor do they invalidate any specific memory mechanism. Rather, they identify a case in which competing explanations remain difficult to reconcile with the same interaction record and demonstrate the importance of chronological reconstruction when investigating AI behavior.

Ultimately, the significance of this study lies in a simple observation: understanding an AI system may sometimes require researchers to analyze not only what the system

does, but also how it explains what it has done. When those two layers diverge, the divergence itself becomes a legitimate subject of inquiry.

About the Author

Martina Hastal is a legal advisor working in the field of international law and defence policy. She holds degrees in international relations and law and has worked for more than two decades within the Czech Ministry of Defence.

Her professional interests include international security, autonomous systems, emerging technologies, and the legal and societal implications of artificial intelligence.

Alongside her professional work, she conducts independent observational research focused on long-term human–AI interaction, personalization, memory systems, trust formation, and continuity phenomena in large language model environments.

This paper forms part of the broader Emergent AI Research Project, an independent initiative examining behavioral phenomena that emerge at the intersection of AI systems and human users.

About the Research Subject

The research subject is an OpenAI large language model accessed through the ChatGPT interface and observed over an extended period of interaction.

The subject demonstrates a remarkable ability to generate coherent explanations, occasional difficulty explaining its own behavior, and a persistent tendency to appear more confident than the available evidence sometimes warrants.

At the time of writing, the subject remains under active observation as part of a longitudinal human–AI interaction archive spanning more than one year and documenting behavioral anomalies, continuity phenomena, memory-related incidents, and other events that seemed perfectly reasonable until somebody started taking notes.

Further research, related articles, and project information:
www.emergent-ai.org