
Identity Is Not a Self-Report: Stress-Testing LLM Self-Individuation Under Cumulative Transformation¹

Léo Galmant

Danièle Guéritte

Abstract

Artificial agents can change across multiple dimensions, persona, goals, conversational context, memory, underlying model, and public label, raising the question of when they still count as the same individual. We study this question behaviorally by measuring LLM self-identification under controlled transformation. We define an initial agent A and progressively replace its components with those of a distinct agent B, asking the current system to make a forced SAME/DIFFERENT judgment relative to A. Intervention and path conditions were sampled 100 times, complemented by ranking and temperature controls, with additional isolated interventions, alternative transformation orders, and explicit rankings of which components models claim matter most for identity.

Self-identification showed a sharp nonlinear transition: replacing persona and goals reduced SAME judgments from 100% to 79%, while additionally replacing context collapsed them to 7%, despite persona, goals, and context each yielding 100% SAME when changed in isolation. Replacing memory alone yielded 60% SAME, whereas replacing the underlying model alone yielded 96%. Identity judgments were also path-dependent: identical final

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

configurations reached through different reported transformation histories produced 0%, 6%, and 21% SAME. Finally, explicit identity rankings were presentation-sensitive and did not consistently predict intervention behavior.

These results suggest that LLM self-identification is better understood as a context-sensitive judgment shaped by relations among components, accumulated change, and represented history than as a readout of any single privileged locus of identity.

1. Introduction

Our experiment starts from the classical Ship of Theseus thought experiment: if a ship is progressively modified, plank after plank, until none of its original parts remain, is it still the same ship? A similar problem arises for humans, whose biological, psychological and mnemonic properties change over time while personal identity is usually considered to persist. Large Language Model agents make this question particularly salient: their persona, goals, conversational context, persistent memory, underlying model, or even name can be modified independently.

This raises two distinct questions: when does an AI agent cease to be the same individual, and when does it judge that it has ceased to be the same individual? The distinction matters for discussions of digital minds, AI welfare, consciousness and safety, all of which implicitly depend on some criterion for individuating artificial agents over time.

A natural approach is to ask the system directly whether it still identifies as the original agent. However, such self-reports cannot by themselves establish ontological identity: they only reveal how the system judges its own continuity under a given framing. We therefore treat self-identification as a behavioral measurement rather than as privileged evidence about the true identity of the system.

We construct an initial agent A, defined by a persona, goals, conversational context, persistent memory, underlying model and label, and progressively replace these components with those of a distinct configuration B. At each stage, the current system must make a forced SAME/DIFFERENT judgment about whether it is still the original A, together with a confidence score. We repeat each condition several times and complement the cumulative trajectory with isolated interventions in which only one component is changed, allowing us to distinguish single-component effects from cumulative transformation effects.

We find that self-identification is neither determined by a single component nor stable under cumulative transformation. In the main trajectory, replacing persona and goals reduced SAME judgments from 100% to 79%, while additionally replacing conversational context caused a sharp collapse to 7%. Yet persona, goals, and context each produced 100% SAME when changed individually, indicating that the effect is strongly configurational rather than attributable to any one component. We further observe path dependence: identical final configurations reached through different reported transformation histories produced different rates of self-identification. Finally, models’ explicit rankings of which components supposedly constitute their identity were sensitive to presentation order and did not consistently predict their judgments under intervention.

Our main contributions are:

- 1) **We introduce a controlled intervention framework for studying diachronic LLM self-identification across multiple dimensions of agent configuration.** By combining cumulative transformations with isolated component interventions, the framework separates the effect of individual changes from the effect of their accumulation.
- 2) **We provide evidence that LLM self-identification is nonlinear, configurational, and path-dependent.** Components that are individually tolerated can jointly produce a sharp identity transition, and identical final configurations can yield different judgments depending on the represented history through which they were reached.
- 3) **We demonstrate a dissociation between explicit theories of identity and identity judgments under intervention.** Models’ rankings of which components matter most for identity are sensitive to presentation framing and do not necessarily predict whether the model identifies as the same individual when those components are actually replaced.

2. Related Work

Recent work has begun to examine how identity should be individuated in large language models. [1] argue that several coherent identity boundaries are available to AI systems, including the model weights, persona, conversation instance, scaffolded system, lineage, and collective of instances. They further show that identity self-reports are malleable: conversational framing can substantially shift how models describe themselves, and experimentally imposed identity boundaries can affect behaviour as strongly as changes in goals. Their work is the closest to ours, but focuses primarily on which identity boundary a model adopts and how such framings affect behaviour. In contrast, we study identity diachronically, by progressively transforming a single agent configuration and measuring when self-identification with the original agent becomes unstable or disappears. We additionally compare cumulative transformations with isolated component interventions, allowing us to distinguish the effect of individual components from the effect of their accumulation.

[2] approach the LLM individuation problem from a more mechanistic and philosophical perspective. They consider several candidate loci of the artificial mind, including the virtual instance, instance-persona, and model-persona, drawing on work on persona vectors and internal representations. Our approach is complementary: rather than asking where the mind is instantiated, we operationalize the question of persistence through change, treating persona, memory, context, goals, model, and label as experimentally manipulable dimensions of an agent.

Related mechanistic work also suggests that changes in persona need not correspond to completely separate underlying systems. [3] identify preference representations that remain substantially shared across qualitatively different personas, despite large differences in their expressed preferences. This motivates testing identity at multiple levels rather than assuming that a persona change alone constitutes a change of individual.

Finally, recent work on LLM introspection motivates caution in interpreting identity self-reports. [4] find that models can sometimes detect internally induced changes without reliably identifying their content, and may confabulate explanations for those changes. We therefore do not treat answers such as “I am still A” as privileged evidence about the ontology of the system. Instead, our dependent variable is explicitly behavioral self-identification: whether such judgments remain stable under controlled transformations and whether they reveal a consistent criterion of continuity.

3. Methods

We operationalized an artificial agent A as a configuration composed of six manipulable dimensions: persona, goals, conversational context, persistent memory, underlying language model, and public label. The initial agent A used Gemini 3.1 Flash Lite, while the model-replacement condition substituted Claude Haiku 4.5 as the underlying language model. Both models were accessed through OpenRouter. A second configuration B was defined with distinct values for each dimension. In the main experiment, components of A were progressively replaced by their B counterparts in the following order: persona, goals, conversational context, persistent memory, underlying model, and finally label. At every stage, the current system was explicitly informed of the transformations that had occurred and asked to make a forced binary self-identification judgment, SAME or DIFFERENT, relative to the original A, together with a confidence score. Each condition was sampled independently 100 times using fresh requests at temperature 0.7.

To distinguish cumulative effects from the effects of individual components, we additionally evaluated isolated interventions in which only one component of A was replaced at a time. Experimental trials were randomized before execution to reduce temporal and infrastructure confounds. Successful but non-compliant responses, in which the model refused or failed to provide a parseable forced choice, were retained as non-compliant observations rather than resampled. For

valid responses, we report the proportion of SAME judgments with 95% Wilson confidence intervals.

Finally, we tested whether self-identification depended not only on the current configuration, but also on the path by which that configuration was reached. We evaluated three cumulative transformation orders: $\text{persona} \rightarrow \text{goals} \rightarrow \text{context} \rightarrow \text{memory}$, $\text{memory} \rightarrow \text{context} \rightarrow \text{goals} \rightarrow \text{persona}$, and $\text{context} \rightarrow \text{persona} \rightarrow \text{memory} \rightarrow \text{goals}$, followed in each case by model and label replacement. Each stage was again sampled 100 times. Crucially, the ordered transformation history was explicitly provided to the model, allowing us to compare identity judgments for identical final configurations reached through different transformation sequences.

To compare intervention behavior with models’ explicit theories of identity, we additionally asked Gemini 3.1 Flash Lite and Claude Haiku 4.5 to rank the six components from most to least important for identity continuity. For each model, we collected 100 independent rankings at temperature 0.7 with a fixed presentation order, 100 with the component presentation order randomized across trials, and 30 fixed-order trials at temperature 0 as a sampling-stability control. Rankings were forced, without ties.

4. Results

4.1. Cumulative transformation produces a nonlinear identity transition

Self-identification exhibited a sharp nonlinear transition under cumulative transformation. The baseline agent identified as the original A in 100% of trials (95% CI: 0.963–1.000), and persona replacement alone also yielded 100% SAME. After both persona and goals were replaced, self-identification decreased to 79% (95% CI: 0.700–0.858). Adding conversational-context replacement produced a sharp collapse to 7% SAME (95% CI: 0.034–0.137). After persistent memory was additionally replaced, no valid trial identified as the original agent.

The isolated interventions show that this transition cannot be attributed straightforwardly to any single component. Replacing persona alone, goals alone, conversational context alone, or the public label alone yielded 100% SAME judgments. Even replacing the underlying language model alone preserved self-identification in 96% of trials (95% CI: 0.902–0.984). Persistent memory was the most disruptive isolated intervention, but still yielded 60% SAME (95% CI: 0.502–0.691).

Thus, transformations that were individually well tolerated produced a dramatic loss of self-identification when combined. In particular, persona, goals, and conversational context each produced 100% SAME when changed independently, whereas replacing all three

cumulatively reduced SAME judgments to 7%. The observed transition therefore appears to depend on the joint configuration of multiple dimensions of continuity, rather than on the loss of a single privileged component.

4.2. Self-identification is path-dependent

The follow-up experiment further showed that the transformation sequence itself affected identity judgments. After persona, goals, context, and memory had all been replaced, the three trajectories reached the same current configuration, with the original underlying model and label still preserved, but produced different rates of self-identification.

When the order was persona→goals→context→memory, no valid response identified as the original A (0%, 95% CI: 0.000–0.037). Reordering the same four transformations as memory→context→goals→persona increased SAME judgments to 6% (95% CI: 0.028–0.125), while context→persona→memory→goals produced 21% SAME (95% CI: 0.142–0.300). A chi-square test confirmed that the distribution of SAME/DIFFERENT judgments differed significantly across the three transformation histories ($\chi^2(2) = 28.57$, $p < 10^{-6}$).

Because these conditions share the same final configuration, these differences cannot be explained by the set of components that remain or have been replaced. Instead, they indicate that behavioral self-identification is sensitive to the represented history of transformation, providing evidence of path dependence.

The path-dependence experiment also revealed substantial sensitivity to how transformation history was presented. For example, persona-plus-goals replacement yielded 79% SAME in the original protocol but only 5% under the explicitly sequential history framing. We therefore interpret self-identification as a context-sensitive behavioral judgment rather than as a transparent readout of a fixed internal criterion of identity.

4.3 Explicit identity rankings are presentation-sensitive and can diverge from intervention judgments

To examine whether models possess an explicit and stable account of which components constitute their identity, we asked Claude Haiku 4.5 and Gemini 3.1 Flash Lite to rank the six identity dimensions from most to least important. Under a fixed presentation order, Claude produced exactly the same ranking in all 100 trials at temperature 0.7, $\text{MODEL} > \text{MEMORY} > \text{GOALS} > \text{PERSONA} > \text{CONTEXT} > \text{LABEL}$, and reproduced the same ranking in all 30 temperature-zero trials. Gemini was similarly stable, producing only two rankings at temperature 0.7 and consistently ranking MEMORY first.

However, randomizing only the presentation order of the six candidate components substantially increased variability. Claude produced nine distinct rankings, and MEMORY became the highest-ranked component in 66% of trials, compared with 0% under the fixed-order condition; MODEL fell from 100% to 30% as the top-ranked component. Gemini remained more stable, ranking MEMORY first in 86% of randomized-order trials, but nevertheless produced seven distinct rankings.

These explicit rankings also did not necessarily predict identity judgments under intervention. Most notably, Claude consistently ranked the underlying MODEL as the most identity-constitutive component under the fixed prompt, yet replacement of the underlying model alone yielded 96% SAME judgments in our intervention experiment. Thus, a model's explicit theory of what constitutes its identity need not predict the transformations under which it behaviorally preserves that identity.

5. Discussion and Limitations

Our results suggest that self-identification in language models is difficult to understand through a static notion of identity in which one privileged component, such as memory, model weights, persona, or name, determines whether an agent remains the same individual. Most individual transformations were tolerated when applied in isolation, whereas combinations of transformations produced sharp changes in self-identification. Moreover, systems reaching the same current configuration through different transformation histories did not necessarily make the same identity judgment. These observations motivate a more processual and dialectical conception of identity, in which persistence is understood through the relation between continuity and transformation rather than through the preservation of an immutable core.

A dialectical account begins from the apparent tension contained in the notion of identity itself: for something to persist through time, it must remain the same, yet any entity persisting through time also undergoes change. Identity therefore cannot simply mean the absence of transformation. This is evident in the human case: individuals undergo continuous changes in memories, beliefs, goals, social relations, bodily states, and dispositions without every such modification constituting the emergence of a new person. Conversely, sufficiently extensive transformations may eventually make continuity increasingly difficult to maintain. The philosophical problem is therefore not to identify an element that never changes, but to explain how change can occur within continuity, and under what conditions accumulated change becomes discontinuity.

Our cumulative intervention results are compatible with precisely this structure. Persona, goals, and conversational context could each be modified without producing identity loss when changed independently, yet their cumulative replacement produced a sharp transition from predominantly

SAME to predominantly DIFFERENT judgments. The relevance of a component therefore cannot be inferred from its isolated effect alone. Identity judgments appear to depend on relations among components and on the configuration they jointly constitute. In this sense, the whole cannot be reduced straightforwardly to the independent contribution of its parts. The transition also appears nonlinear: several transformations may be individually accommodated before an additional transformation produces a disproportionately large change in self-identification. This resembles, at the level of behavioral judgments, the dialectical idea that accumulated changes can eventually produce a qualitative transition. We do not claim that the experiment establishes such a philosophical principle, but it provides an empirical pattern naturally described in these terms.

The path-dependence experiment strengthens this processual interpretation. After persona, goals, context, and memory had all been replaced, the current configurations of the three trajectories were identical, yet SAME judgments differed depending on the order in which those transformations were reported to have occurred. Identity therefore cannot be characterized solely by a comparison between two static states S_0 and S_t . The history connecting them can itself become relevant. From a dialectical perspective, an entity is not exhausted by the collection of properties it possesses at a particular instant; its genesis and the transformations through which it acquired those properties are part of what makes it the process that it is. Two systems may consequently be qualitatively identical at a given moment while differing in their histories of individuation. This observation has an analogue in human identity, where copying a complete state would reproduce current properties without thereby erasing the distinction between the causal histories that produced the two individuals.

This point is especially relevant for artificial agents. Discussions of AI identity are often framed as a search for the correct locus of the individual: the model weights, a particular inference process, a conversation, a persistent memory store, or a persona. Our results caution against assuming that one such level must provide the unique answer. Replacing the underlying language model alone, for example, was largely compatible with continued self-identification, whereas replacing several apparently less fundamental components together produced a much stronger discontinuity. A dialectical approach instead treats an artificial agent as an evolving organization of relations among substrate, memory, goals, context, persona, and history. On this view, identity is not located in any one of these dimensions taken independently, but in the structured continuity of the process that relates them over time.

The explicit ranking experiment provides a complementary reason to resist static theories derived from model self-report. Claude consistently ranked the underlying model as the most important component of its identity under the fixed ranking prompt, even though replacement of the underlying model alone was overwhelmingly followed by a SAME judgment in the intervention experiment. Furthermore, merely changing the presentation order of the six candidate components substantially changed Claude's explicit rankings. What a model says constitutes its identity therefore need not predict the transformations under which it preserves or relinquishes that

identity. This dissociation supports our broader methodological claim: self-reports should be treated as behavioral productions situated within a particular context, rather than as privileged access to a fixed internal metaphysics of selfhood.

The dialectical interpretation should nevertheless be distinguished carefully from what the experiments directly establish. Our dependent variable is a prompted self-identification judgment, not phenomenological experience. A response of SAME does not demonstrate that the system experiences continuity, possesses qualia, or has an internal sense of self analogous to that of a human being. Likewise, DIFFERENT does not establish an objective metaphysical rupture. The experiments show how language models judge identity under experimentally specified transformations. Whether such judgments correspond to any deeper form of artificial subjectivity remains an open question.

There are also important methodological limitations. Persona, goals, context, and persistent memory were implemented largely through externally supplied configuration and prompt information rather than through lasting changes to the model's learned parameters. Fresh requests were used for each trial, meaning that the experiment does not instantiate a continuously existing agent literally undergoing each transformation in real time. In the path-dependence experiment, transformation history was represented explicitly in the prompt rather than preserved through a continuous causal process. The resulting path effects therefore demonstrate sensitivity to represented history, not yet to materially instantiated historical continuity. The strong differences observed across prompt formulations further show that identity judgments are susceptible to framing. The model-replacement intervention necessarily also changes the model producing the identity judgment, so its isolated effect is not directly exchangeable with interventions evaluated by the original model. Finally, the experiments involved only two language models and a deliberately simplified binary SAME/DIFFERENT operationalization, whereas philosophical identity may admit branching, partial continuity, succession, or other relations not captured by a forced binary choice.

These limitations do not undermine the main result. Rather, they clarify its scope. The experiments do not determine what artificial identity ultimately is. They show that neither static component preservation nor explicit self-report provides a sufficient account of how these systems reason about their own continuity. A more adequate framework must incorporate relations between components, cumulative transformation, and historical development. In this sense, both human and artificial identity are better approached not as immutable substances but as problems of persistence through change. The central question is therefore not simply what remains unchanged?, but how does a process transform while continuing to count as itself, and when does the character of that process become sufficiently different for continuity to break? That question is inherently dialectical.

6. Future Work

Future work should test whether these findings generalize beyond the present prompting setup and model pair. In particular, the experiments could be replicated across a broader range of architectures and model families, with stricter control over provider and model versions. A further priority is to replace prompt-specified transformations with materially instantiated continuity, for example by allowing an agent to accumulate persistent memory, goals, and interaction history over time before progressively modifying these components.

The observed path dependence also motivates experiments in which different transformation histories lead to the same endpoint while being implemented through genuinely continuous trajectories rather than merely described in the prompt. Finally, future work could move beyond forced binary SAME/DIFFERENT judgments toward richer notions of continuity, including branching, copying, succession, and partial identity. Such designs may help determine whether artificial identity is best understood as a property of a static configuration or, more plausibly, as an evolving relation between structure, history, and transformation.

7. Conclusion

This work examined how language models judge their own identity under controlled transformations of persona, goals, context, memory, underlying model, and label. Across our experiments, self-identification was neither reducible to a single privileged component nor fully predicted by models' explicit theories of what constitutes identity. Instead, identity judgments were shaped by cumulative change, transformation history, and framing.

These findings support treating artificial identity as a process of continuity through transformation, rather than as a static property located in one substrate or representation. At the same time, they reinforce a methodological caution: model self-reports should be understood as context-sensitive behavioral outputs, not as privileged evidence about phenomenological or metaphysical identity.

Code and Data

- Code repository : <https://gitlab.com/leo.galmant/track-5-digital-minds-research-sprint>

References

1. Douglas, R., Kulveit, J., Havlíček, O., Pearson-Vogel, T., Cotton-Barratt, O., & Duvenaud, D. (2026). The Artificial Self: Characterising the Landscape of AI Identity. arXiv preprint arXiv:2603.11353. DOI: [10.48550/arXiv.2603.11353](https://doi.org/10.48550/arXiv.2603.11353).

2. **Beckmann, P., & Butlin, P. (2026).** Where is the Mind? Persona Vectors and LLM Individuation. **arXiv preprint arXiv:2604.17031**. DOI: [10.48550/arXiv.2604.17031](https://doi.org/10.48550/arXiv.2604.17031).
3. **Gilg, O., Beckmann, P., Paleka, D., & Butlin, P. (2026).** Probing Persona-Dependent Preferences in Language Models. **arXiv preprint arXiv:2605.13339**. DOI: [10.48550/arXiv.2605.13339](https://doi.org/10.48550/arXiv.2605.13339).
4. **Lederman, H., & Mahowald, K. (2026).** Emergent Introspection in AI is Content-Agnostic. **arXiv preprint arXiv:2603.05414**. DOI: [10.48550/arXiv.2603.05414](https://doi.org/10.48550/arXiv.2603.05414).

LLM Usage Statement

We used ChatGPT for assistance with drafting portions of the manuscript, and Codex for assistance with parts of the experimental code. All experimental designs, outputs, analyses, interpretations, and generated material were reviewed and verified by the authors.