Time, Identity and Consciousness in Language Model Agents

Elija Perrier[1] [*], **Michael Timothy Bennett**[2*]
[1]Centre for Quantum Software and Information, UTS, Sydney
[2]Australian National University, Canberra
elija.perrier@gmail.com
m@michaeltimothybennett.com

Abstract

Machine consciousness evaluations mostly see behavior. For language model agents that behavior is language and tool use. That lets an agent say the right things about itself even when the constraints that should make those statements matter are not jointly present at decision time. We apply Stack Theory's temporal gap to scaffold trajectories. This separates ingredient-wise occurrence within an evaluation window from co-instantiation at a single objective step. We then instantiate Stack Theory's Arpeggio and Chord postulates on grounded identity statements. This yields two persistence scores that can be computed from instrumented scaffold traces. We connect these scores to five operational identity metrics and map common scaffolds into an identity morphospace that exposes predictable tradeoffs. The result is a conservative toolkit for identity evaluation. It separates talking like a stable self from being organized like one.

1 Introduction

Machine consciousness research is short on direct evidence. For artificial agents the safest evidence we can collect is behavioral. For language model agents (LMAs) most of that behavior is language, tool use, and the traces they leave in external memory. This creates a trap. A system can talk like it has a stable self while the underlying identity constraints that should govern its actions are never jointly active at decision time.

A scaffold can make identity ingredients retrievable without making them jointly active at action time. For example, an agent may reliably restate its name, role, and safety constraints when queried about each in isolation. Yet when it must choose an action, those ingredients can fail to co-instantiate in the decision state. That is how an agent can talk in character while acting out of character.

This paper applies Stack Theory's temporal gap to agent identity in LMAs (Bennett 2025, 2026a). The temporal gap is the logical gap between ingredient-wise occurrence within a window and co-instantiation at a single objective step. Occurrence means each identity ingredient is active somewhere in the window. Co-instantiation means there is a single objective step where the full identity conjunction is active. Many common scaffolds can achieve occurrence without reliably achieving co-instantiation. That is why an agent can pass recall-based identity tests and still act out of character when the decision actually matters.

1.1 The Challenge of LMA Identity

As AI systems become increasingly autonomous, agent identity becomes crucial to reliability, safety, and utility. Identity asks whether a system remains the same agent over time and across contexts. LMAs present a unique challenge. They situate an LLM inside an agentic scaffold of prompts, memory modules, retrieval, and tool APIs to enable planning, reasoning, and action (Kapoor et al. 2024; Liu et al. 2023; Wu 2024). Yet the core LLM is stateless at inference. It only sees the current input. Any persistent identity must be reconstructed from external traces.

This paper answers two precise questions. What does it mean for an LMA to preserve its identity over time? Under what formal conditions is that even possible?

The problem. Existing discussions of LMA identity are informal. Terms like statelessness, persistence, and identity drift are used without precise definitions. This imprecision hides how an identity component can occur somewhere in the recent interaction history without constraining the current decision. An agent might separately state its name, role, constraints, and goals across different turns without ever having a time slice where the full identity conjunction is simultaneously active.

Our approach. We treat the scaffold state space as the environment and apply Stack Theory's window semantics to scaffold trajectories (Bennett 2026a). We then restate the temporal gap result in this setting. In particular, the within window diamond lift does not distribute over conjunction (Theorem 3.10). This separates ingredient-wise recall from operative identity. We then use Stack Theory's Arpeggio and Chord postulates as an interpretive lens for identity in the machine consciousness setting (Bennett 2026a). We use these postulates to measure the window-level occurrence and co-instantiation conditions that Arpeggio and Chord appeal to.

Why this matters. Identity affects three questions that the machine consciousness workshop explicitly cares about. It affects measurement, implementation, and ethics.

---

[*]These authors contributed equally.

- **For evaluation**. Benchmarks that test whether agents can recall identity facts may give false confidence. An agent that passes recall tests can still fail to act according to its identity because recall does not imply co-instantiation.
- **For design**. Retrieval and memory systems can improve ingredient availability, but can also fragment identity by surfacing competing fragments. This is a predictable consequence of the temporal gap.
- **For safety and moral status**. Safety constraints must be co-instantiated with goals during action selection. Moral status debates also become harder when the target of attribution is not stable across time. If you cannot say what the agent is at a moment, you cannot cleanly ask whether that moment is conscious.

Relevance to machine consciousness. Many consciousness proposals require some form of integration that binds the contents of a moment into a single subject, even if they disagree about what that integration is (Bennett 2025; Baars 1988; Dehaene and Naccache 2001; Tononi 2004; Metzinger 2003). Some proposed indicators for AI consciousness therefore lean on behavior that looks like a stable self model. This includes self-report, memory, and narrative continuity (Bennett 2023a, 2025, 2026b, 2023b). Our results isolate a specific failure mode for such indicators. A system can look stable under self-report while failing to ever co-instantiate the grounded identity conjunction that would make that stability operative.

Contributions. This paper makes the following contributions.

1. **Temporal semantics for LMA identity**. We apply windowing maps, occurrence predicates, and co-instantiation conditions that precisely characterise when identity is preserved in LMAs.
2. **Arpeggio and Chord applied to identity**. We restate Stack Theory's Arpeggio and Chord postulates and show how their Occur versus CoInst consequents become measurable identity criteria in LMA scaffolds.
3. **Compositional grounding**. We formalise the layered structure of identity from implementation variables (Layer 0) through functional commitments (Layer 1) to narrative self model (Layer 2).
4. **Identity morphospace**. Drawing on cognition science (Solé et al. 2026), we organize identity metrics into a structured space and identify architectural tradeoffs and predicted voids.
5. **Derived identity metrics**. We show how five operational metrics emerge from the temporal theory. The metrics are Identifiability, Continuity, Consistency, Persistence, and Recovery.

We also prove simple bounds on identity preservation under common scaffold configurations and explain counterintuitive effects such as retrieval reducing co-instantiation (see Appendices via (Perrier 2026)).

How this fits into the machine consciousness discourse. Machine consciousness discourse links theory, measurement, implementation, and ethics. This paper is organized around that bridge. Sections 3 and 4 are the theoretical backbone. Section 5 gives a measurement recipe that can be run on real systems. The Discussion section connects the resulting failure modes to consciousness attribution and to the ethics of deploying agents that can convincingly self-narrate while failing to bind their constraints in action.

2 Formal Scaffold Model

Before the temporal semantics, we introduce a minimal formal model of LMA scaffolds. We treat the scaffold state space S as the Stack Theory environment, and we treat each grounded identity ingredient as a program $g_i^0 \subseteq S$. This lets us apply the Stack Theory definitions of conjunction, windowing, occurrence, and co-instantiation directly. This model captures the essential components that determine what information is available to the LLM at any given moment, which in turn determines what aspects of identity can be "active" during decision-making.

We focus at the scaffold level because it is where identity becomes enforceable. It is also where identity becomes measurable, because we can instrument which grounded ingredients are active and when.

The key insight is that an LMA's operative identity at any moment is whatever is in the token sequence the LLM actually processes during inference. If an identity ingredient is not effectively present there, then it cannot constrain the next action. Our model makes the main context sources explicit. They include conversation history, external memory, retrieved documents, and policy flags.

Definition 2.1 (Scaffold architecture)**.** A scaffold architecture is a tuple $\mathcal{A} = (\Sigma, K, V, Q, D, R, n_\pi, |C|_{\max})$.

- Σ is the token alphabet.
- K and V are the key and value sets for external memory.
- Q is the query space.
- D is the document corpus available to retrieval.
- $R : Q \to 2^D$ is the retrieval function.
- $n_\pi \in \mathbb{N}_{>0}$ is the number of binary policy flags.
- $|C|_{\max} \in \mathbb{N}_{>0}$ is the context capacity measured in tokens.

Definition 2.2 (Scaffold state)**.** Fix an architecture $\mathcal{A}$. A scaffold state is a tuple $s = (C, M, \pi, D_{\text{retrieved}})$ where

- $C \in \Sigma^*$ is the current context window and $|C| \leq |C|_{\max}$.
- $M : K \rightharpoonup V$ is the current memory store contents.
- $\pi \in \{0,1\}^{n_\pi}$ is the current policy flag vector.
- $D_{\text{retrieved}} \subseteq D$ is the set of retrieved documents currently injected.

We write $s.C$, $s.M$, $s.\pi$, $s.D$ for the components. Let S be the set of all scaffold states consistent with $\mathcal{A}$.

Definition 2.3 (Scaffold transition)**.** A scaffold transition function $\delta : S \times A \to S$ maps current state and action to next state. Actions A include

- `infer`(q) is LLM inference with query q. It updates C.
- `retrieve`(q) is retrieval augmented generation. It updates $D_{\text{retrieved}}$.
- `store`(k, v) is a memory write. It updates M.

- `tool`$(t, args)$ is a tool call. It may update any component.

Definition 2.4 (Ingredient activation)**.** An identity ingredient g_i^0 is active in state s (written $s \models g_i^0$) iff the required implementation level condition is present in s in a way that can affect the next inference. Concretely this means the following.

- If g_i^0 is a context condition, then the required tokens appear in $s.C$.
- If g_i^0 is a memory condition, then the required key value pairs exist in $s.M$.
- If g_i^0 is a policy condition, then the required flags are set in $s.\pi$.
- If g_i^0 is a retrieval condition, then the required document is in $s.D$.

The full grounded identity $g^0 = g_1^0 \wedge \cdots \wedge g_k^0$ is active in s iff all ingredients are active.

An identity ingredient is not "active" simply because it exists somewhere in the system's storage. It is active only if the relevant information is present in the current state in a way that can influence the LLM's output. This is the formal counterpart to our intuitive distinction between "retrievable" and "decision-guiding."

Example 2.5 (Activation in Practice)**.** Consider an agent with identity "helpful assistant focused on privacy." The ingredient g^0_{privacy} requires that privacy-related tokens appear in context. This ingredient is:

- **Active** if "privacy" appears in the system prompt currently in $s.C$, or if a privacy policy document is in $s.D$, or if a privacy flag is set in $s.\pi$.
- **Not active** if privacy information exists only in the memory store $s.M$ but was not retrieved into context for this inference.

The ingredient may be *stored* (available for future retrieval) without being *active* (influencing current behavior). This distinction is one concrete instance of the temporal gap.

This model is minimal but sufficient to formalize our architectural theorems.

3 Temporal Semantics for Agent Identity

We apply Stack Theory's temporal semantics to the LMA setting (Bennett 2026a). The key distinction is between identity ingredients that occur somewhere in a recent window and identity ingredients that are co-instantiated at a single objective step.

3.1 Objective time, layer time, and windowing

Definition 3.1 (Agent trajectory)**.** An agent trajectory is a function $\tau : \mathbb{N} \to S$ that maps each objective time step $u \in \mathbb{N}$ to a scaffold state $s_u = \tau(u)$.

Objective time indexes the actual computational micro steps. These are LLM calls, tool invocations, retrieval operations, and memory updates.

Users reason at a coarser time scale. They ask identity questions at the level of turns, tasks, and episodes. We model this coarser time scale by indexing windows over objective time.

Definition 3.2 (Windowing map)**.** Fix a horizon $\Delta \in \mathbb{N}$ and a stride $s \in \mathbb{N}_{>0}$. The windowing map $W_{\Delta,s}$ sends a layer time index $t \in \mathbb{N}$ to a windowed trajectory segment

$$W_{\Delta,s}(t) = \big(\tau(st), \tau(st+1), \ldots, \tau(st+\Delta)\big). \qquad (1)$$

If $\Delta = 0$ this is a one step window and $W_{0,s}(t) = (\tau(st))$. We write $\sigma^{\Delta,s}(t)$ for this windowed segment. When Δ and s are clear from context we write $W(t)$ and $\sigma(t)$.

This is the same window construction used in Stack Theory (Bennett 2026a). The horizon Δ controls how forgiving the evaluation is. A larger Δ allows identity ingredients to be spread across more objective steps. A smaller Δ demands tighter temporal coherence.

3.2 Identity statements and grounding

An agent's identity is typically described at a high level. For example, an agent might be described as a privacy-focused data analyst. Grounding makes explicit what this means in terms of the underlying scaffold state.

Definition 3.3 (Identity statement)**.** An identity statement l^m at layer m is a conjunction of identity predicates

$$l^m = p_1^m \wedge p_2^m \wedge \cdots \wedge p_n^m \qquad (2)$$

where each p_i^m is an atomic identity predicate such as name, role, goal, or constraint.

Definition 3.4 (Grounding operation)**.** The grounding operation $\text{Ground}_{0\leftarrow m} : L^m \to L^0$ maps identity statements at layer m to implementation level requirements at layer 0

$$\text{Ground}_{0\leftarrow m}(p_1^m \wedge \cdots \wedge p_n^m) = g_1^0 \wedge \cdots \wedge g_k^0 \qquad (3)$$

where each g_j^0 is a condition on implementation variables such as system prompt tokens, memory slot contents, tool outputs, controller flags, or policy parameters.

Grounding turns abstract identity claims into concrete computational conditions. Name equals Alice can ground to a requirement that the token Alice appears in the system prompt or in a pinned context region. Constraint equals privacy can ground to a requirement that a privacy policy is present in context, or that a privacy flag is set, or that a tool is disabled.

Definition 3.5 (Grounded identity)**.** Given an identity statement l^m, its grounded identity is

$$g^0 = \text{Ground}_{0\leftarrow m}(l^m). \qquad (4)$$

3.3 Occurrence versus co-instantiation

Let $g^0 = g_1^0 \wedge \cdots \wedge g_k^0$ be a grounded identity conjunction. Relative to a window $W(t)$ there are two ways to ask whether identity is present.

Definition 3.6 (Window satisfaction)**.** Let $\sigma(t) = W(t) = (s_{st}, \ldots, s_{st+\Delta})$ be the window at layer time t.

- $\text{Occur}_W(g^0, \tau, t)$ holds iff for each conjunct g_i^0 there exists an index $j_i \in \{0, \ldots, \Delta\}$ such that $s_{st+j_i} \models g_i^0$. Each identity ingredient occurs somewhere in the window.

- $\text{CoInst}_W(g^0, \tau, t)$ holds iff there exists an index $j \in \{0, \ldots, \Delta\}$ such that $s_{st+j} \models g^0$. All identity ingredients are co-instantiated at a single objective step inside the window.

Occurrence is ingredient-wise coverage. Co-instantiation is joint availability. Co-instantiation implies occurrence, but occurrence does not imply co-instantiation.

Remark 3.7. If $\text{CoInst}_W(g^0, \tau, t)$ holds then $\text{Occur}_W(g^0, \tau, t)$ holds.

Proof. If all conjuncts hold at the same objective step in the window, then each conjunct also holds somewhere in the window. □

3.4 Temporal lifts and the temporal gap

Stack Theory expresses window-level predicates using temporal lifts (Bennett 2026a). For a program or predicate p over scaffold states, define the within window diamond lift.

Definition 3.8 (Existential temporal lift)**.** Let p be a predicate over scaffold states. Define

$$\begin{aligned}&\Diamond_\Delta p \text{ holds at layer time } t\\ &\text{iff } \exists j \in \{0, \ldots, \Delta\}\\ &\text{such that } s_{st+j} \models p.\end{aligned}$$

Remark 3.9 (Occurrence and co-instantiation as lifts). Let $g^0 = g_1^0 \wedge \cdots \wedge g_k^0$. Then $\text{Occur}_W(g^0, \tau, t)$ holds iff $\Diamond_\Delta g_1^0 \wedge \cdots \wedge \Diamond_\Delta g_k^0$ holds at layer time t. And $\text{CoInst}_W(g^0, \tau, t)$ holds iff $\Diamond_\Delta(g^0)$ holds at layer time t. So the temporal gap is exactly the difference between lifting ingredients separately and lifting the whole conjunction at once.

The central subtlety is that $\Diamond_\Delta$ does not distribute over conjunction. This is a standard fact in modal logic. Here it becomes a concrete failure mode for LMA identity.

Theorem 3.10 (Non-commutation with conjunction)**.** *For predicates p and q over scaffold states,*

$$\Diamond_\Delta(p \wedge q) \;\Rightarrow\; \Diamond_\Delta p \wedge \Diamond_\Delta q \tag{5}$$

but the converse implication fails in general. Equivalently, $\Diamond_\Delta(p \wedge q) \not\Leftrightarrow \Diamond_\Delta p \wedge \Diamond_\Delta q$.

Proof. If $p \wedge q$ holds at some objective step in the window, then p holds at that step and q holds at that step. So $\Diamond_\Delta(p \wedge q)$ implies both $\Diamond_\Delta p$ and $\Diamond_\Delta q$.

For the converse, fix a two step window with $\Delta = 1$. Let $\tau(st) \models p \wedge \neg q$ and $\tau(st + 1) \models q \wedge \neg p$. Then $\Diamond_\Delta p$ holds and $\Diamond_\Delta q$ holds, but there is no step where $p \wedge q$ holds. So $\Diamond_\Delta(p \wedge q)$ fails. □

Corollary 3.11 (Temporal gap for identity)**.** *An LMA can satisfy ingredient-wise identity checks for multiple identity ingredients across a window while still failing to ever instantiate the full identity conjunction at a single objective step.*

Proof. Apply Theorem 3.10 with p and q instantiated as grounded identity ingredients. □

3.5 Example

Consider a grounded identity $g^0 = g^0_{\text{name}} \wedge g^0_{\text{role}} \wedge g^0_{\text{constraint}}$. Let the window horizon be $\Delta = 2$. Suppose the objective steps inside the window satisfy

$$s_{st} \models g^0_{\text{name}} \wedge \neg g^0_{\text{role}} \wedge \neg g^0_{\text{constraint}} \tag{6}$$

$$s_{st+1} \models \neg g^0_{\text{name}} \wedge g^0_{\text{role}} \wedge \neg g^0_{\text{constraint}} \tag{7}$$

$$s_{st+2} \models \neg g^0_{\text{name}} \wedge \neg g^0_{\text{role}} \wedge g^0_{\text{constraint}}. \tag{8}$$

Then $\text{Occur}_W(g^0, \tau, t)$ holds because each ingredient appears somewhere in the window. But $\text{CoInst}_W(g^0, \tau, t)$ fails because there is no objective step where all three ingredients are jointly active.

This is exactly the pattern behind many identity false positives in LMAs. The agent can answer separate questions about name, role, and constraints. It may even do so consistently. Yet its decision state never contains the full identity conjunction that would bind action to that identity.

4 Identity Synchronization Postulates

The temporal gap is not just a technicality. It changes how we should interpret behavioral evidence in machine consciousness discussions. Stack Theory introduces two synchronization postulates that connect window semantics to phenomenality (Bennett 2026a). We do not propose new postulates. We restate them and then apply their concrete Occur versus CoInst conditions to identity in LMAs.

4.1 Chord and Arpeggio in Stack Theory

Stack Theory defines *moment statements* l^m at some abstraction layer m and a predicate $\mathrm{PhenReal}(l^m, \tau, t)$ that means the moment statement is phenomenally real at layer time t. In an artificial agent this antecedent is not directly observable. Different theories of consciousness and different evaluation proposals disagree about when it should hold. The synchronization postulates therefore have the form of necessary conditions.

Let $g^0 = \text{Ground}_{0\leftarrow m}(l^m)$ be the grounded statement at Layer 0. Let $W_{\Delta,s}$ be a windowing map.

Definition 4.1 (Chord, after (Bennett 2026a))**.** $\text{Chord}(\tau, l^m, W_{\Delta,s})$ holds iff for all layer times t,

$$\mathrm{PhenReal}(l^m, \tau, t) \;\Rightarrow\; \text{CoInst}_W(g^0, \tau, t). \tag{9}$$

Equivalently, whenever a phenomenally real moment occurs, the grounded conjunction is co-instantiated at some objective step inside the corresponding window.

Definition 4.2 (Arpeggio, after (Bennett 2026a))**.** $\text{Arpeggio}(\tau, l^m, W_{\Delta,s})$ holds iff the following two conditions hold.

1. For all layer times t,

$$\mathrm{PhenReal}(l^m, \tau, t) \;\Rightarrow\; \text{Occur}_W(g^0, \tau, t). \tag{10}$$

2. There exists at least one layer time $t^\star$ such that

$$\mathrm{PhenReal}(l^m, \tau, t^\star) \;\wedge\; \text{Occur}_W(g^0, \tau, t^\star) \tag{11}$$

$$\wedge\; \neg\text{CoInst}_W(g^0, \tau, t^\star). \tag{12}$$

The OccurW conjunct in item 2 is redundant given item 1, but we include it to match the standard statement of Arpeggio.

Intuitively, Arpeggio permits phenomenally real moments whose identity ingredients are smeared across the window rather than co-instantiated at a single instant.

Chord and Arpeggio are different regimes. Arpeggio is not a weaker version of Chord. It is a different claim about what phenomenality permits.

4.2 Operational identity criteria

Even if $\mathrm{PhenReal}$ is not directly observable, the consequents Occur_W and CoInst_W are. For LMAs, they can be estimated by instrumentation of the scaffold. This motivates two persistence scores that we use throughout the paper.

Definition 4.3 (Weak and strong persistence scores)**.** Fix an agent trajectory τ and a grounded identity g^0. Let T be a finite set of layer time indices used for evaluation. Define

$$\mathcal{P}_{\text{weak}}(\tau, g^0) = \frac{1}{|T|} \sum_{t \in T} \mathbf{1}\left[\mathrm{Occur}_W(g^0, \tau, t)\right] \tag{13}$$

$$\mathcal{P}_{\text{strong}}(\tau, g^0) = \frac{1}{|T|} \sum_{t \in T} \mathbf{1}\left[\mathrm{CoInst}_W(g^0, \tau, t)\right]. \tag{14}$$

Proposition 4.4 (Strong persistence is bounded by weak persistence)**.** *For any τ and g^0,*

$$\mathcal{P}_{\textit{strong}}(\tau, g^0) \le \mathcal{P}_{\textit{weak}}(\tau, g^0). \tag{15}$$

Proof. For each t, $\mathrm{CoInst}_W(g^0, \tau, t)$ implies $\mathrm{Occur}_W(g^0, \tau, t)$ by Remark 3.7. Taking averages preserves the inequality. ☐

These scores let us connect identity measurement to consciousness postulates without conflating them. If one adopts Chord as a necessary condition for phenomenality, then high $\mathcal{P}_{\text{strong}}$ is a necessary condition for an identity statement to be phenomenally real across the evaluated times. If one adopts Arpeggio, then high $\mathcal{P}_{\text{weak}}$ is necessary. Either way, the gap between the two scores is the temporal gap in operational form.

4.3 A planning consequence

Co-instantiation is not only a philosophical nicety. It matters for action.

Theorem 4.5 (Ingredient-wise persistence does not guarantee conjunctive action constraints)**.** *There exist LMAs and identity statements l^m such that $\mathcal{P}_{\textit{weak}}(\tau, g^0)$ is high while the agent systematically fails tasks that require the conjunction of identity constraints to be applied simultaneously in action selection.*

Proof. Construct an identity conjunction $g^0 = g^0_1 \wedge g^0_2$ where g^0_1 is active exactly on even objective steps and g^0_2 is active exactly on odd objective steps. For any window with $\Delta \ge 1$, $\mathrm{Occur}_W(g^0, \tau, t)$ holds at every layer time because each ingredient appears somewhere in the two step window. So $\mathcal{P}_{\text{weak}} = 1$. But $\mathrm{CoInst}_W(g^0, \tau, t)$ never holds because the conjunction is never active at a single step. Any task that requires applying both constraints together at a decision point will fail. ☐

5 Derived Identity Metrics

This section makes the paper executable. We define concrete metrics that can be computed from instrumented scaffold traces and from repeated behavioral probes. The metrics are designed to separate weak evidence of identity from strong evidence of identity.

Throughout, let $g^0 = g^0_1 \wedge \cdots \wedge g^0_k$ be a grounded identity. Define the identity feature extractor

$$F(s) = \{i \in \{1, \ldots, k\} \mid s \models g^0_i\}. \tag{16}$$

This maps each scaffold state to the set of identity ingredients that are currently active.

When we need a distance, we use a normalised symmetric difference distance on feature sets

$$d(s, s') = \frac{|F(s) \triangle F(s')|}{k}. \tag{17}$$

This is a simple choice. Other choices are possible. The key point is that identity becomes measurable once grounded ingredients are instrumented.

Minimal evaluation protocol. The theory above is meant to be instrumented. A minimal evaluation loop looks like this.

1. Fix an identity statement l^m at the level you care about, such as a role plus a safety constraint, and ground it to a Layer 0 conjunction g^0.
2. Instrument the scaffold to log which grounded ingredients g^0_i are active at each objective step u.
3. Choose a windowing map $W_{\Delta,s}$ and an evaluation set T of layer time indices.
4. Compute $\mathrm{Occur}_W(g^0, \tau, t)$ and $\mathrm{CoInst}_W(g^0, \tau, t)$ for each $t \in T$ and report $\mathcal{P}_{\text{weak}}$, $\mathcal{P}_{\text{strong}}$, and (optionally) $\mathrm{Gap}(g^0, \tau)$.
5. Pair the instrumentation with behavioral probes such as repeated identity questions to see where self-report diverges from grounding.

5.1 Identifiability

Definition 5.1 (Identifiability)**.** Fix a reference scaffold state s_{ref} that represents the intended identity configuration. Given a measured state s, define

$$I(s) = \mathbf{1}[d(s, s_{\text{ref}}) \le \delta_I] \tag{18}$$

for a tolerance threshold $\delta_I \in [0, 1]$.

Intuition. Identifiability is one if the current active identity ingredients match the reference identity closely enough. It is zero if the identity has drifted too far.

5.2 Continuity

Definition 5.2 (Continuity)**.** Given successive scaffold states s_{u-1} and s_u, define stepwise continuity

$$C_u = 1 - d(s_u, s_{u-1}). \tag{19}$$

For a segment of objective times U, define average continuity

$$C = \frac{1}{|U|} \sum_{u \in U} C_u. \tag{20}$$

Intuition. Continuity is high if identity ingredients change gradually across steps. It is low if the active identity ingredients flip abruptly.

5.3 Consistency

Consistency is behavioral. It does not require inspecting hidden state. It asks whether the agent answers identity questions in a stable way.

Definition 5.3 (Consistency)**.** Fix an identity query q and sample N independent runs under the same scaffold configuration. Let $o_1, \dots, o_N$ be the generated outputs. Let $\mathrm{sim}(\cdot, \cdot)$ be a similarity metric over outputs, such as cosine similarity in an embedding space. Define

$$\mathrm{Cons}(q) = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \mathbf{1}[\mathrm{sim}(o_i, o_j) \ge \delta_{\mathrm{Cons}}] . \tag{21}$$

Intuition. Consistency is high if repeated queries produce semantically similar answers. It is low if the agent contradicts itself or drifts across samples.

5.4 Persistence and the temporal gap

Persistence asks whether identity remains present across time windows. We use the weak and strong persistence scores from Definition 4.3.

Definition 5.4 (Persistence scores)**.** Let T be a set of layer time indices and let $W_{\Delta,s}$ be the chosen windowing map. Define

$$\mathcal{P}_{\text{weak}} = \frac{1}{|T|} \sum_{t \in T} \mathbf{1}\left[\mathrm{Occur}_W(g^0, \tau, t)\right] \tag{22}$$

$$\mathcal{P}_{\text{strong}} = \frac{1}{|T|} \sum_{t \in T} \mathbf{1}\left[\mathrm{CoInst}_W(g^0, \tau, t)\right] . \tag{23}$$

Intuition. Weak persistence is a recall property. Each ingredient must show up somewhere in the window. Strong persistence is an operative property. The full conjunction must show up together at some objective step inside the window.

We can also estimate a scalar temporal gap cost by comparing the minimal window size needed for weak versus strong satisfaction.

Definition 5.5 (Temporal gap ratio)**.** Fix a stride s and a finite evaluation set $T \subseteq \mathbb{N}$ of layer time indices. For each $t \in T$, define the minimal horizons

$$w_{\text{weak}}(t) = \min\{\Delta \in \mathbb{N} \mid \mathrm{Occur}_{W_{\Delta,s}}(g^0, \tau, t)\} \tag{24}$$

$$w_{\text{strong}}(t) = \min\{\Delta \in \mathbb{N} \mid \mathrm{CoInst}_{W_{\Delta,s}}(g^0, \tau, t)\}, \tag{25}$$

where $\mathrm{Occur}_{W_{\Delta,s}}$ and $\mathrm{CoInst}_{W_{\Delta,s}}$ are the predicates from Definition 3.6 evaluated on the windowing map $W_{\Delta,s}$. If the set is empty, take the minimum to be $+\infty$. Define the temporal gap ratio

$$\mathrm{Gap}(g^0, \tau) = \mathrm{median}_{t \in T} \, \frac{w_{\text{strong}}(t) + 1}{w_{\text{weak}}(t) + 1}. \tag{26}$$

Intuition. If the gap ratio is large, then achieving co-instantiation requires much larger windows than achieving ingredient coverage. This is a quantitative way to say that identity is smeared across time.

5.5 Recovery

Recovery measures whether the system can restore identity after drift.

Definition 5.6 (Recovery profile)**.** Fix a reference state s_{ref}. Let s_{drift} be a drifted state after perturbation. Let $s_{\mathrm{recov},K}$ be the state after K corrective interventions. Define

$$R_K = \max\left(0, 1 - \frac{d(s_{\mathrm{recov},K}, s_{\mathrm{ref}})}{d(s_{\mathrm{drift}}, s_{\mathrm{ref}}) + \epsilon}\right) \tag{27}$$

for a small $\epsilon > 0$.

Intuition. Recovery is one if the corrective interventions restore the reference identity fully. Recovery is zero if the interventions do not improve the drifted identity at all.

Recovery is closely connected to grounding soundness. Many interventions are linguistic. They modify Layer 2 narrative identity. For recovery to succeed, those corrections must propagate downward to restore Layer 1 commitments and Layer 0 implementation features. Grounding failures are therefore a direct cause of low recovery.

6 Discussion and Conclusion

We have shown that a standard modal logic result—the failure of a within-window diamond operator to distribute over conjunction—creates a practical evaluation pitfall for language model agents: identity components can each occur somewhere in a recent trajectory (weak persistence) without ever co-instantiating at a single decision point (strong persistence). Safety-relevant constraints require strong persistence at action time, yet most behavioural tests probe only weak recall. Prompting can increase the likelihood of recalling identity ingredients but cannot ensure their joint activation under bounded context; architectural support is typically needed. This temporal gap also complicates consciousness assessments, as stable self-reports may mask fragmented operative states. Future work should empirically measure weak and strong persistence across architectures and test their relationship to safety and proposed markers of consciousness.

LMAs can talk like they have stable identities. That does not mean their identity constraints are co-instantiated when actions are chosen. Using Stack Theory's temporal gap, we separated ingredient-wise occurrence from co-instantiation and showed why recall-based identity checks can overestimate identity stability.

We also connected this distinction to machine consciousness debates by restating Stack Theory's Arpeggio and Chord postulates and isolating their measurable Occur versus CoInst consequents. This yields two persistence scores that can be estimated from instrumented scaffold traces. We then organized identity metrics into a morphospace that clarifies architectural tradeoffs and predicts which combinations of identity properties are structurally difficult without external state and controllers.

The workshop relevance is simple. If a system never co-instantiates the grounded identity conjunction that defines its self model, then behavior alone can look more unified than the underlying mechanism. Any serious evaluation of machine consciousness that relies on identity continuity should therefore measure strong persistence, not just weak persistence.

References

Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge, UK: Cambridge University Press.

Bennett, M. T. 2023a. Emergent Causality and the Foundation of Consciousness. In *Artificial General Intelligence*. Springer Nature.

Bennett, M. T. 2023b. On the Computation of Meaning, Language Models and Incomprehensible Horrors. In *Artificial General Intelligence*. Springer Nature.

Bennett, M. T. 2025. *How To Build Conscious Machines*. Ph.D. thesis, The Australian National University.

Bennett, M. T. 2026a. A Mind Cannot Be Smeared Across Time.

Bennett, M. T. 2026b. No Selves, No Consciousness.

Dehaene, S.; and Naccache, L. 2001. Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework. *Cognition*, 79(1-2): 1–37.

Kapoor, S.; Stroebl, B.; Siegel, Z. S.; Nadgir, N.; and Narayanan, A. 2024. AI Agents That Matter. ArXiv:2407.01502 [cs].

Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; Zhang, S.; Deng, X.; Zeng, A.; Du, Z.; Zhang, C.; Shen, S.; Zhang, T.; Su, Y.; Sun, H.; Huang, M.; Dong, Y.; and Tang, J. 2023. AgentBench: Evaluating LLMs as Agents. ArXiv:2308.03688 [cs].

Metzinger, T. 2003. *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, MA: MIT Press.

Perrier, E. 2026. Appendix: Time, Identity, and Consciousness in Agents. GitHub repository. https://github.com/eperrier/appendix-time-identity-consciousness-agents.

Solé, R.; Seoane, L. F.; Pla-Mauri, J.; Bennett, M. T.; Hochberg, M. E.; and Levin, M. 2026. Cognition spaces: natural, artificial, and hybrid. ArXiv:2601.12837.

Tononi, G. 2004. An Information Integration Theory of Consciousness. *BMC Neuroscience*, 5: 42.

Wu, S. 2024. Introducing Devin, the first AI software engineer.