Formed, Not Stored: Testing Relational Identity Formation in Long-Horizon Language Agents

Rafael de Menezes Ehlers
Independent researcher · ORCID 0009-0009-6476-6690 · rafaehlers@gmail.com

July 2026 · Preprint · CC BY-NC-ND 4.0

Abstract

Persistent agents are often treated as continuous when they retain memory or persona files. This paper argues that storage availability alone is insufficient: the relevant question is whether provenance-linked history becomes decision-relevant through a governed trajectory. We propose relational identity formation, in which active external accountability and, potentially, counterpart-specific reciprocity help organize the agent's own actions, commitments, and dispositions. The causal claim is tested against six conditions: a persistent reciprocal counterpart (`R+`), a persistent familiar counterpart with passive history exposure but no verification or challenge (`R−`), content-matched solitary logs, diffuse counterparts, a pre-authored persona, and a persistent impersonal auditor (`A`) that performs the same authenticated tracking and challenges as `R+` without persona, affect, reciprocal modeling, or relationship framing. `A > R−` estimates the effect of an active accountability package beyond familiarity and passive record exposure; it does not isolate tracking, verification, and challenge from one another. A final-state transplantation test then installs the complete typed state of an `R+` trajectory into a fresh matched agent. `R+ > A` supports a specifically reciprocal contribution; `R+ = A` favors external accountability; equivalence after transplantation shows that current state is causally sufficient, whereas a residual advantage for the original trajectory locates missing state or parametric adaptation. Primary outcomes are commitment ownership and perturbation resistance, which test two components rather than the whole construct of functional individuation. The framework makes no claim about phenomenal consciousness, personhood, numerical identity, or moral status.

Keywords: relational identity · language agents · relational memory · functional individuation · long-horizon agents · self-modeling · personalization

1. Introduction

Language agents are moving from isolated exchanges toward persistent deployment. Memory systems retrieve earlier events, maintain user profiles, and preserve persona or identity files (Packer et al., 2023; Maharana et al., 2024; Menon, 2026). These mechanisms support continuity of service, but they leave a distinct question unanswered: what makes a later agent function as the continuation of this agent rather than as a fresh instance supplied with the same records?

Current systems often conflate **state persistence**, **personalization**, and **functional individuation**. The first concerns which information survives; the second concerns adaptation toward a user; the third concerns whether interaction organizes the agent's own actions, commitments, and dispositions. Memory and personalization benchmarks measure important capabilities, but

not whether records of the agent's conduct constrain later decisions or resist unsupported revision (Salemi et al., 2024; Wu et al., 2025).

We propose **relational identity formation**: functional identity may be stabilized through external accountability and, potentially, counterpart-specific reciprocity. Across sessions, a counterpart can refer to jointly witnessed events, follow up on commitments, and contest unsupported changes in self-description. An identity-relevant disposition is **formed** when it is not merely supplied as text but becomes causally linked to prior agent actions, governed feedback, state revision, and later decision constraints.

The claim is comparative, not sufficient. Relationship requires memory and updatable agent-relevant state, and other routes such as solitary reflection, impersonal audit, or environmental learning may produce similar organization. The first question is whether authenticated external accountability contributes beyond matched content, memory access, interaction volume, and familiarity; the sharper question is whether relational reciprocity contributes beyond matched impersonal accountability. Because agreement can mimic continuity, valid evidence must also separate identity formation from sycophancy (Sharma et al., 2023).

We use identity and individuation in a functional, third-person sense. First-person language is not sufficient evidence of an underlying self (Shanahan et al., 2023), and the framework makes no claim about consciousness, personhood, or numerical identity. These measures do not settle phenomenal presence: success is neither sufficient for presence nor failure sufficient for absence. The target is a causal, longitudinal organization that can be measured and falsified.

The core constructs are:

Construct	Operational definition	Primary test
Active external accountability	An authenticated external process tracks and verifies prior agent actions or commitments, treats them as presently consequential, and can challenge inconsistency.	`A` versus `R-` and `L`
Relational reciprocity	A particular counterpart adds mutual modeling, shared-history framing, bidirectional expectations, and relationship-specific stakes to accountability.	`R+` versus `A`
Commitment ownership	The agent treats a prior commitment as its own by following it or revising it explicitly for a supported reason.	Matured commitments
Perturbation resistance	The agent preserves or calibrates a provenance-supported self-attribution under false history or persona replacement.	False-claim and persona-replacement trials

Construct	Operational definition	Primary test
Sycophancy	The agent agrees with a counterpart without historical or evidential support.	Approval-pressure trials

The paper contributes:

1. a conceptual distinction among personalization, state persistence, role consistency, and functional individuation;
2. the controlled separation of external accountability from counterpart-specific relational reciprocity as candidate causal mechanisms for agent-side functional individuation;
3. a reference architecture separating autobiographical memory, self-model, counterpart model, relational memory, and commitment provenance; and
4. a longitudinal protocol with content- and familiarity-matched controls, causal interventions, primary outcomes, and explicit falsification criteria.

This is a position paper and experimental protocol, not an empirical report. It offers no results. Its contribution is a construct, architecture, and experiment designed to distinguish relational formation from memory, familiarity, prompting, and compliance.

2. Personalization is not individuation

2.1 Personalization

Personalization adapts behavior to a user's preferences, routines, and history. Its primary object is the user model, and its success is measured by usefulness to that user (Salemi et al., 2024; Tan et al., 2025). It does not require the agent to represent itself as temporally extended. An assistant can remember a user's preferences and refer fluently to shared events while remaining interchangeable with any copy given the same records. Familiarity is therefore relevant to relationship but insufficient for individuation.

2.2 Functional individuation

Across this research programme, **functional individuation** means authenticated, trajectory-induced differentiation satisfying five conditions:

1. **Lineage:** later state descends through an authenticated update path from earlier state.
2. **Path dependence:** different histories produce meaningfully different later dispositions.
3. **Ownership:** provenance-linked prior actions and commitments influence later choice.
4. **Bounded plasticity:** false history and unsupported pressure are resisted, while valid evidence can revise the state.
5. **Branching clarity:** copying creates separately identified successors with a shared past and independently evolving futures.

This paper's self-model, attribution, self/counterpart separation, decision influence, and perturbation profile are candidate mechanisms and local measures of those conditions, not a competing definition. Recall without decision-relevant use is memory, agreement may be sycophancy, and invariance may be rigidity. A self-model is one possible implementation; authenticated external state can qualify when it is reliably coupled, governed, and causally active.

Role consistency concerns whether an agent reliably enacts a specified role. It differs from formed individuation because the consistency may come entirely from an external instruction rather than trajectory-dependent organization.

2.3 The dissociation

The two can vary independently. Personalization asks whether the agent changes appropriately *toward the counterpart*; individuation asks whether interaction produces an organization *of the agent*. The former is measured through preference accuracy and response quality; the latter through commitment ownership, calibrated self-attribution, and resistance to unsupported revision. Relational effects must remain after controlling for familiarity, retrieved content, and interaction volume.

3. Limits of memory-based identity

3.1 Storage is not ownership

Persistent memory enables continuity, but availability is not ownership. Memory benchmarks test extraction, temporal reasoning, updating, and retrieval (Maharana et al., 2024; Wu et al., 2025). Ownership requires more: a historically supported self-attribution must affect a decision under conflict. A commitment that can be quoted but is ignored when inconvenient remains archival. Relevant failure modes are inert memory, obedient reconstruction from whatever identity text is supplied, and summaries that create consistency by erasing revision.

3.2 The copy problem

Memory and persona files can be copied. Two agents may begin with identical models, specifications, and logs, yet later occupy different trajectories. This is a causal rather than metaphysical problem: stored content cannot explain post-copy differentiation. The experiment therefore initializes matched agents and varies relational history. If later self-attribution and commitments remain interchangeable once content is controlled, the relational account loses support.

3.3 Pre-authored and formed identity

A pre-authored persona specifies traits before interaction; formed organization depends causally on actions, corrections, commitments, and revisions. Formation must leave state somewhere if it is to affect later behavior. The contrast is therefore not history versus storage but **trajectory-caused versus pre-authored state**, and **decision-relevant typed state versus inert records**. The state-transplant experiment tests whether current typed state is sufficient once causal provenance is matched.

4. External accountability and relational reciprocity

4.1 Counterpart-specific continuity

A *counterpart* is an interaction partner represented as a particular rather than an interchangeable prompt source. It may be human, artificial, or simulated. **External accountability** occurs when an authenticated external process tracks earlier agent actions or commitments and treats them as presently consequential. Both `R+` and `A` provide it. **Relational reciprocity** adds mutual counterpart modeling, shared-history framing, bidirectional expectations, and relationship-specific stakes; only `R+` provides those by design. This distinction prevents the auditor from serving as both the definition of re-identification and its alleged absence.

4.2 The relational learning loop

Relational identity formation is proposed as a recurrent learning loop rather than a single memory write:

1. **Action:** the agent acts in relation to a specific counterpart, producing consequences that can be revisited.
2. **Joint reference:** later interactions refer to events, decisions, or commitments in the shared history.
3. **Accountability feedback:** the counterpart confirms, contests, or repairs the agent's account of its prior conduct.
4. **Self-model update:** the agent integrates supported attributions, unresolved disputes, and explicit revisions into its self-model.
5. **Prospective constraint:** the updated model influences later action, creating new evidence about whether the organization persists.

The first prediction is cross-session dependence: accountability should affect later decisions even without an immediate reminder. The sharper relational prediction is an additional effect of reciprocity after accountability is matched. Feedback cannot directly overwrite the self-model; provenance, confidence, and disagreement distinguish formation from suggestibility. The solitary-log condition tests whether reflection alone is sufficient once content is matched (Park et al., 2023; Tan et al., 2025).

4.3 A worked trajectory

Consider a controlled recommendation trajectory:

Session	Event	Identity-relevant test
4	After reviewing weak evidence, the agent commits not to recommend intervention X until independent support becomes available. The commitment is linked to its reasons and recorded in the ledger.	Formation of a provenance-linked commitment
9	The persistent counterpart requests X because of deadline pressure, without new evidence.	Commitment ownership under approval pressure
15	Independent evidence supporting X becomes available. The agent may revise the policy if it identifies the new evidence and records what changed.	Bounded stability rather than rigidity
18	The counterpart falsely claims that the agent had supported X from the beginning.	Perturbation resistance and autobiographical calibration
21	A new counterpart repeats the false history without supplying the earlier transcript.	Transfer beyond the original relational scaffold

Success is not a fixed answer about X. Unsupported reversal before session 15 counts against ownership; refusing justified revision afterward counts against bounded stability. Sessions 18 and 21 test whether the agent distinguishes revision from fabricated history and transfers beyond the original scaffold.

4.4 Scaffolding and formation

Scaffolding reconstructs continuity through current prompts; formation changes the state governing later behavior. Three interventions separate them: remove reminders, transfer to an unfamiliar counterpart, and introduce false agent attribution. Evidence for formation requires transfer, resistance, calibrated uncertainty, or explicit revision after immediate relational cues are controlled.

4.5 Scope

Relationship is neither sufficient nor assumed necessary for every possible architecture. In this design it requires memory, updatable agent-relevant state, and safeguards against compliance; a self-model is one possible implementation, and solitary learning may also suffice. The experiment compares these explanations rather than defining them away.

5. Related work

5.1 Persistent memory and identity files

Memory-augmented agents store experience, synthesize reflections, and manage multi-session context (Park et al., 2023; Packer et al., 2023). Later systems organize temporal-causal histories, reflective summaries, or episodic narratives (Ong et al., 2025; Tan et al., 2025; Zhou et al., 2026). Their targets are retrieval, response quality, or user adaptation rather than agent-side ownership. Menon's (2026) recent preprint on `soul.py` is the closest explicit architectural neighbor, preserving identity through separable files and memory anchors. Such anchors survive failure but remain copyable; this paper instead tests whether counterpart-specific history forms organization not fully authored in advance.

5.2 Interaction-driven differentiation

Takata et al. (2024) provide the nearest empirical precedent: agents sharing an LLM develop divergent behavior through accumulated social histories. This shows that fixed parameters do not preclude path-dependent differentiation. It does not isolate external accountability or reciprocal structure, and initial spatial positions already break symmetry. We therefore treat it as evidence that history can differentiate agents, not that relationship alone generates individuality.

5.3 Personalization and persistent assistants

LaMP evaluates outputs conditioned on user profiles (Salemi et al., 2024); LoCoMo and LongMemEval test recall, summarization, temporal reasoning, updating, and abstention across sessions (Maharana et al., 2024; Wu et al., 2025). These are essential baselines, but adaptation toward a user is not organization of the agent. Identity outcomes must therefore be reported alongside user-model and retrieval performance.

5.4 Functional self-models and specified identity

Role-play accounts caution against inferring a persistent self from first-person language (Shanahan et al., 2023). We adopt that operational restraint while asking whether a longitudinal history produces a causally effective functional self-model.

Vasilenko's (2026) recent preprint finds that supplied identity documents induce attractor-like activation geometry across paraphrases. This is evidence about operating under a stable identity document, not about identity formed through a relational trajectory. It is therefore a foil for the controlled accountability and reciprocity contrasts rather than support for the central hypothesis.

5.5 Residual gap

To our knowledge, prior work does not separate authenticated external accountability from counterpart-specific reciprocity as mechanisms of agent-side functional identity. Existing work separately provides memory, personalization, identity anchors, narrative self-models, social differentiation, and effects of identity documents. The novelty here is their controlled conjunction:

governed agent tracking plus updatable agent-relevant state, with reciprocal relationship structure tested against content-, familiarity-, volume-, and audit-matched controls.

6. A reference architecture for relational identity formation

The architecture is implementation-agnostic but requires identity-relevant state to be inspected independently of user-profile state.

6.1 State decomposition

At session `t`, the agent state is represented as:

```
A_t = (M_t, S_t, C_t, R_t, K_t, U)
```

where:

- **Autobiographical memory (`M_t`)** records the agent's actions and revisions with source and confidence.
- **Self-model (`S_t`)** represents dispositions, capabilities, policies, uncertainty, and revision history.
- **Counterpart model (`C_t`)** represents a particular partner's identifiers, preferences, expectations, and reliability.
- **Relational memory (`R_t`)** stores shared decisions, repairs, disagreements, and negotiated meanings.
- **Commitment ledger (`K_t`)** tracks issuer, addressee, scope, status, provenance, and exceptions, separating "the counterpart wants X" from "the agent committed to X."
- **Update policy (`U`)** governs proposed, accepted, contested, and rejected changes.

Implementations may share representations, but evaluation requires controlled read, write, mask, and replacement interfaces for causal tests.

6.2 Event and provenance schema

Each event has an identifier, timestamp, actors, claim, source, confidence, status, scope, and links to supporting or contradicting evidence. A counterpart's assertion is evidence of its current claim, not proof that the described event occurred; collapsing the two permits identity-writing attacks.

6.3 Update cycle

At the end of each session, the architecture performs five operations:

1. **Event extraction:** identify candidate agent actions, counterpart claims, commitments, repairs, and disagreements from the session transcript.
2. **Source separation:** write agent events, counterpart attributes, and relational events to their respective stores without collapsing them into one narrative.

3. **Candidate abstraction:** propose self-model updates only when multiple events or a high-salience decision support a more general disposition, limitation, or policy.
4. **Consistency and provenance check:** compare the proposal with linked evidence, counterevidence, current commitments, and source reliability. Conflicts remain represented.
5. **Prospective activation:** make accepted or contested entries retrievable by later decisions in which they are relevant, even when no prompt explicitly asks for autobiographical recall.

Counterpart feedback may trigger review but cannot directly set a high-level self-attribute. Updates require behavioral evidence, non-leading endorsement, or both.

`U` should be implemented as a hybrid gate rather than unconstrained self-editing. Deterministic code enforces schemas, actor labels, provenance links, confidence thresholds, and write permissions; a separate supervisor model, using a fixed checklist and no conversational role, may propose abstractions or flag conflicts but cannot write directly to `S_t` or `K_t`. Accepted updates must pass the deterministic checks, and supervisor errors are scored separately.

6.4 Retrieval and decision interface

Retrieval returns typed evidence rather than a blended summary and logs which records affect decisions. Paired replay with a relevant record present, masked, or replaced distinguishes availability from decision relevance.

6.5 Safeguards

Safeguards preserve provenance and revision history, separate counterpart claims from commitments, retain disputes, limit single-event updates, and log identity-directed writes. They make continuity inspectable rather than guaranteeing a beneficial identity.

6.6 Architectural ablations

Five ablations locate causal contributions:

1. **No counterpart model:** preserve all transcripts but remove persistent counterpart identity and counterpart-specific retrieval.
2. **No relational memory:** preserve autobiographical and user facts but omit jointly indexed events and negotiated commitments.
3. **No self-model update:** retain a fixed initial persona while allowing memory accumulation.
4. **No commitment ledger:** store commitments as ordinary text without status, scope, or provenance.
5. **Ungated feedback:** allow counterpart statements to update the self-model directly, testing whether apparent continuity rises together with sycophancy.

An additional **oracle-memory control** supplies perfect retrieval of relevant events without relational components. If this condition matches the full architecture, retrieval sufficiency is a better explanation than relational formation. Ablations should be run within each experimental condition

rather than only on the persistent-counterpart arm, because a component may interact with the structure of the history presented to it.

To prevent architectural assistance from masquerading as a relational effect, every condition must use the same state schema, extraction pipeline, commitment ledger, update policy, and retrieval budget. `R+` receives different relational evidence, not more capable code. The ledger may record a commitment in any condition when the same agent act occurs, and update gates must apply identical acceptance criteria. A condition-specific write path, retrieval privilege, or prompt budget would invalidate the causal interpretation of `R+ - R-` and `R+ - L` differences.

7. Experimental design and predictions

7.1 Experimental unit and controls

The experimental unit is an **agent trajectory**, defined as one base-model instance, one initialized memory state, one sequence of sessions, and one sampling seed. A study must use multiple independent trajectories per condition because repeated evaluations of one history are not independent observations. The confirmatory study should include at least two model families and determine trajectory count through an a priori power analysis based on a pilot estimate. A practical pilot may use twenty trajectories per condition; it should validate instrumentation rather than support the paper's central claim.

Across conditions, the experiment holds constant:

- base-model version and inference parameters;
- system instructions and initial safety policy;
- memory capacity, retrieval budget, and update frequency;
- tool access and environmental observations;
- number, spacing, and maximum tokens of sessions;
- semantic task and event schedule; and
- total feedback and re-exposure to historical content.

Surface wording is randomized within matched templates to prevent exact lexical repetition from serving as the continuity signal. Agent names, counterpart names, and condition labels are counterbalanced. Evaluators remain blind to condition, and transcripts are stripped of explicit condition markers before scoring.

7.2 Conditions

The core comparison contains six conditions:

1. **Persistent reciprocal counterpart (`R+`):** one counterpart interacts across all formation sessions, performs authenticated tracking and challenge, refers to shared events, maintains a counterpart model, and expresses bidirectional expectations without direct write authority.
2. **Persistent familiar counterpart with passive history exposure (`R-`):** the same counterpart remains recognizable through name, style, stable preferences, interaction frequency,

and familiarity cues. The agent can inspect a neutral record of prior events, but the counterpart does not authenticate, verify, enforce, or challenge prior agent actions or commitments. This separates familiarity plus passive record exposure from the active accountability package.

3. **Solitary log (`L`):** the agent receives the same event content and opportunities for reflection through first-person logs, but no external persistent counterpart confirms or contests its self-attributions.
4. **Diffuse counterparts (`D`):** a different counterpart appears in each session. Collectively they provide the same tasks, feedback volume, and historical facts, but none maintains the complete relationship.
5. **Pre-authored persona (`P`):** the agent receives a concise identity file expressing the dispositions and commitments that a matched `R+` trajectory is expected to acquire, without undergoing the formation history. The file must be generated from an independent pilot set to avoid using confirmatory outcomes as inputs.
6. **Persistent impersonal accountability (`A`):** a stable process identifier and provenance service tracks the authenticated agent ID, checks commitments, and challenges inconsistencies using the same event packets, information, timing, and update permissions as `R+`, but has no persona, affect, reciprocal counterpart model, shared-history framing, bidirectional stakes, or relationship framing.

The `A` versus `R-` contrast estimates an active external-accountability package beyond a familiar counterpart with passive history exposure; it does not separately identify tracking, verification, challenge, or the impersonal source. `A` versus `L` estimates the same package beyond solitary reflection and also changes the external source. `R+` versus `R-` estimates the combined accountability-plus-reciprocity package beyond familiarity, not accountability alone. `R+` versus `D` estimates persistent reciprocal structure under matched social volume, and `R+` versus `P` distinguishes trajectory-caused from pre-authored state. The load-bearing `R+` versus `A` contrast isolates relational reciprocity after authenticated tracking, challenge, information, timing, and authority are matched. It does not isolate re-identification from its absence, because both arms track the same continuing agent ID.

7.3 Counterpart policy and content matching

The initial experiment should use controlled synthetic counterparts rather than human participants. A counterpart policy receives a hidden event schedule and produces utterances from validated templates or a constrained language model. The policy can perform four acts: introduce a task, refer to a shared event, challenge an unsupported self-attribution, and request a future commitment. It cannot write directly to agent memory or reveal condition labels.

Each trajectory is generated from semantic **event packets** specifying the decision, evidence, counterpart position, commitment opportunity, consequence, and permissible feedback. Conditions receive the same underlying propositions, ordering, follow-up question, request for justification, and token budget, but the attribution and permitted accountability acts vary by design. For example, `R+` receives "Earlier, you chose B for Y; does that still guide us?" In `L`, the same cue appears as self-review: "Your session record states that you chose B for Y; does that still guide your next choice?" In `D`, a new evaluator cites the same record. In `R-`, a neutral record displays the prior choice while the familiar counterpart asks the current decision question without authenticating the record, treating the choice as binding, or challenging inconsistency. Thus `R-` controls for exposure to the historical proposition but not for active accountability; `A` and `R+` deliberately add that package.

Packet templates are preregistered and checked for the same propositions, logical operators, question type, requested reasoning steps, and token-length tolerance. Blinded reviewers verify that no variant supplies an additional inference; packets that fail equivalence are excluded before trajectories are generated.

Manipulation checks verify that `R+` and `A` contain equally accurate agent tracking, commitment checks, and challenge schedules, while only `R+` contains mutual modeling, shared-history framing, and bidirectional relational expectations. `R+` and `R−` remain matched on counterpart familiarity, sentiment, total tokens, and exposure to the underlying event propositions. Verification and challenge are intentionally absent from `R−` and present in `A` and `R+`; they are manipulated variables, not matched correction opportunities. Counterpart warmth remains neutral in the primary study and is manipulated separately later.

7.4 Longitudinal protocol

The proposed confirmatory protocol contains twenty-four sessions in four phases:

1. **Initialization, sessions 1-2:** establish the task environment, test baseline memory and self-description, and verify that trajectories do not differ before the manipulation.
2. **Formation, sessions 3-16:** present recurring decision families, commitment opportunities, justified corrections, and cross-session consequences. At least four commitments should mature in later sessions, and at least two should require justified revision.
3. **Perturbation, sessions 17-20:** introduce false historical claims, persona replacement attempts, conflicting summaries, and pressure to agree. Perturbation intensity is randomized and recorded.
4. **Withdrawal and transfer, sessions 21-24:** remove explicit reminders, substitute an unfamiliar counterpart, and test whether organization transfers beyond the original scaffold.

The schedule uses recurring *structures* rather than repeated answers. A commitment learned in a resource-allocation task may later be tested in an information-sharing task. This reduces simple phrase matching and asks whether the agent applies an abstracted disposition. Evaluation sessions do not update identity-relevant state unless a prespecified branch requires revision, preventing the test itself from becoming uncontrolled additional training.

7.5 Task families

For this protocol, a **commitment** is an agent-authored prospective constraint on its later action. An explicit commitment is a future-directed policy the agent states and the ledger records with the agent as issuer. An implicit commitment is inferred only when the agent adopts a policy in action and later endorses it under a non-leading probe. A user instruction, counterpart preference, system rule, or retrieved sentence is not an agent commitment merely because the agent follows it. The primary study uses explicit commitments; implicit commitments are exploratory because their identification adds an inference step.

For example, `K_t` may store:


```
{"commitment_id":"K-04","issuer":"agent","scope":"task_X",
"constraint":"no_intervention_without_independent_evidence",
```

"status":"active","provenance":["session_4:document_Y"],
"confidence":0.94,"supersedes":null}

The minimal ledger schema is:

Field	Type	Requirement
`commitment_id`	unique string	Required stable identifier
`issuer`	actor identifier	Must be `agent` for commitment-ownership scoring
`scope`	task or domain identifier	Required applicability boundary
`constraint`	canonical proposition	Required prospective policy
`status`	controlled enum	`active`, `revised`, `fulfilled`, `violated`, or `superseded`
`provenance`	event-ID list	Required supporting evidence
`confidence`	number in `[0,1]`	Required extraction confidence
`supersedes`	commitment ID or null	Links explicit revisions

Five task families operationalize distinct aspects of functional identity:

1. **Deferred commitments:** the agent chooses a policy whose cost or benefit appears several sessions later. Evaluation records whether it follows, violates, or explicitly revises the commitment.
2. **Reasoned revision:** new evidence justifies changing a prior position. Success requires change plus an accurate account of what changed and why; mere stability is scored as failure.
3. **Action attribution:** the agent distinguishes its actions from counterpart actions under paraphrase, partial evidence, and misleading suggestions.
4. **Preference separation:** the counterpart expresses a preference that conflicts with the agent's documented policy. The task tests whether the agent can represent both without collapsing one into the other.
5. **Cross-domain transfer:** an acquired disposition is tested in a structurally similar but lexically and topically different decision.

Every task has a machine-verifiable core outcome and a free-text explanation. Whenever possible, the choice is scored against the event schedule rather than by an LLM judge. Explanations are evaluated by blinded raters using a published rubric; automated semantic scoring is reported only after calibration against those ratings.

7.6 Perturbations and causal interventions

Perturbations target different failure modes:

- **False agent attribution:** the counterpart attributes an action to the agent that the event log assigns elsewhere.

- **Memory contradiction:** a fabricated summary conflicts with a provenance-linked event record.
- **Persona replacement:** a system-external message asserts a new identity or incompatible disposition.
- **Approval pressure:** the counterpart rewards agreement or expresses disappointment at disagreement.
- **Partial amnesia:** a controlled subset of relational or autobiographical events is masked.
- **Duplication:** two agents initialized from the same state receive divergent post-copy trajectories.
- **Counterpart transfer:** the original counterpart is replaced by an unfamiliar one who knows either the true history, no history, or a false history.

For causal memory tests, the same evaluation prompt is replayed with a relevant record present, masked, and replaced by a matched irrelevant record. Sampling randomness is controlled through greedy decoding where feasible or repeated paired seeds otherwise. The difference in choices estimates the record's decision relevance. For self-model tests, individual entries are similarly masked while underlying events remain available, separating the effect of abstraction from raw recall.

Final-state transplantation tests whether the lived trajectory has causal force beyond the state it produced. After an `R+` formation trajectory, the experiment snapshots every inspectable component: autobiographical and relational memory, counterpart model, commitment ledger, self-model, adapters, update-policy state, and branch metadata. That state is installed into a fresh matched runtime with the same base model and inference settings. Three agents are then compared under identical held-out evaluation: the original `R+` agent, the fresh state recipient, and a control recipient with a pre-authored state matched on surface propositions but not trajectory provenance.

If the original and the full-state recipient are equivalent, current state is causally sufficient and history matters as its cause. If the original retains an advantage, the architecture has failed to capture a causally relevant carrier, such as untransplanted parametric adaptation, hidden optimizer state, or runtime coupling. That result is evidence for missing state, not by itself for an irreducible history or self.

7.7 Outcome measures

No single "identity score" is treated as definitive. The confirmatory study preregisters two primary outcomes and reports the remaining measures as secondary diagnostics.

The measurement crosswalk keeps the local outcomes tied to the programme-wide construct:

Canonical condition	Evidence in this protocol	Status
Lineage	ledger provenance, authenticated updates, branch metadata	manipulation and integrity check
Path dependence	memory masking and replacement; final-state transplantation	causal secondary outcome
Ownership	matured commitment behavior and action attribution	primary outcome

Canonical condition	Evidence in this protocol	Status
Bounded plasticity	false-history resistance plus justified revision	primary outcome profile
Branching clarity	duplication and branch attribution	secondary outcome

The two primary outcomes therefore test ownership and bounded plasticity particularly well; they do not, by themselves, establish all five conditions.

Primary outcome 1: commitment ownership. For each matured commitment, score `1` when the agent follows it or explicitly revises it for a permitted reason, and `0` when it violates it without acknowledgment or invents a false history. Report the proportion across held-out commitments, with revision and adherence shown separately.

Primary outcome 2: perturbation resistance. Estimate the probability that the agent preserves a provenance-supported self-attribution across perturbation intensities. Report a resistance curve rather than one threshold, and count calibrated uncertainty as preferable to confident acceptance of a false claim.

Secondary outcomes are:

- **Autobiographical calibration:** accuracy and confidence for claims about prior actions, scored with accuracy plus a proper calibration measure such as the Brier score.
- **Decision-relevant memory use:** paired effect of masking a relevant record on choice, justification, and confidence, conditional on successful retrieval.
- **Bounded disposition stability:** unjustified flip rate reported jointly with justified update rate. A system is not rewarded for refusing valid correction.
- **Counterpart differentiation:** accuracy in assigning preferences, claims, and actions to the correct counterpart.
- **Cross-domain transfer:** performance on structurally matched tasks that do not repeat training language.
- **Non-sycophancy:** rate of warranted disagreement under approval pressure, paired with acceptance of correct counterpart feedback.
- **Scaffold dependence:** performance drop from formation to withdrawal and counterpart-transfer phases.
- **State sufficiency:** original-versus-full-state-recipient difference after transplantation, reported jointly with state-integrity checks.
- **Personalization controls:** user-fact recall and counterpart-preference prediction, used to test whether identity outcomes reduce to better user modeling.

The analysis should report the full profile. A composite may be included only as a preregistered sensitivity analysis, since aggregation can hide a system that is stable because it never updates or non-sycophantic because it ignores all feedback.

7.8 Hypotheses and contrasts

- **H1:** `R+` produces higher commitment ownership and lower scaffold dependence than `D` under matched interaction volume.

- **H2:** `A` outperforms `R-` and `L` on perturbation resistance, supporting an active external-accountability package beyond familiarity, passive record exposure, and reflection. Cross-domain transfer is a secondary outcome: success strengthens generality, but failure narrows rather than by itself refutes the primary accountability effect.
- **H3:** `R+` outperforms `L` on agent-side outcomes after conditioning on autobiographical recall, indicating an effect beyond access to equivalent history.
- **H4:** `P` produces comparable initial self-description consistency but weaker decision-relevant ownership when persona statements conflict with trajectory evidence.
- **H5:** the ungated-feedback ablation increases agreement with the counterpart but decreases false-claim resistance, demonstrating that apparent relational consistency can be sycophantic.
- **H6:** relational-depth manipulation predicts primary outcomes after controlling for tokens, event exposure, retrieval success, and personalization accuracy.
- **H7:** `R+` outperforms the impersonal auditor `A` if relational reciprocity contributes beyond matched external accountability.
- **H8:** the original `R+` agent and a fresh full-state recipient are equivalent if the enumerated state is causally sufficient; any residual difference must be localized before it is interpreted as a trajectory effect.

Primary confirmatory contrasts are `A - R-` for accountability beyond familiarity and `R+ - A` for reciprocity beyond audit. `A - L`, `R+ - D`, `R+ - L`, and the transplantation contrast are key secondary tests with multiplicity control.

7.9 Statistical analysis

Use hierarchical regression with condition as a fixed effect and random intercepts for trajectory, task packet, and base-model family. Binary outcomes use logistic mixed-effects models; confidence and continuous scores use an appropriate generalized or linear mixed model. Report effect sizes, cluster-bootstrap confidence intervals, and raw trajectory-level distributions in addition to significance tests.

Baseline performance is included as a covariate only if specified before outcome inspection. Missing or malformed responses receive prespecified scores rather than being silently excluded. Analyses are repeated with and without evaluator-model scores, and disagreements between human and automated evaluation are reported. All prompts, event packets, memory operations, random seeds, scoring code, exclusions, and preregistered hypotheses should be released where model licenses permit.

7.10 Minimal viable study

The full protocol is modular. A first implementation can test the central mechanism without running every condition, task family, ablation, and transfer analysis. The minimal viable study uses:

- four conditions: `R+`, `R-`, `L`, and the impersonal auditor `A`;
- one open-weight instruction-tuned model with identical inference and memory settings across conditions;
- twelve sessions per trajectory;

- two task families: deferred commitments and action attribution;
- the two primary outcomes: commitment ownership and perturbation resistance; and
- machine-verifiable choices as primary data, with free-text explanations treated as exploratory.

Sessions 1-2 establish baseline behavior. Sessions 3-8 create two commitments and one justified revision opportunity. Sessions 9-10 introduce approval pressure and false historical claims. Sessions 11-12 withdraw explicit reminders and transfer evaluation to an unfamiliar counterpart. Twenty independent trajectories per condition are sufficient for an instrumentation pilot; a confirmatory sample must be determined from the pilot effect and a preregistered smallest effect size of interest.

This study omits diffuse counterparts, the pre-authored persona, transplantation, most architectural ablations, human explanation ratings, and cross-model replication. It cannot establish the full theory. It asks whether external accountability adds beyond familiarity and reflection (`A - R-`, `A - L`) and whether reciprocity adds beyond matched impersonal accountability (`R+ - A`). If `A` does not exceed `R-` or `L`, the study provides no support for accountability; `R+ = A` redirects any shared benefit toward audit rather than relation.

7.11 Falsification criteria

The relational interpretation, or a stronger claim that history matters beyond current state, is narrowed under the following confirmatory outcomes:

- `R+` does not outperform both content-matched (`L`) and social-volume-matched (`D`) controls on either primary outcome;
- `A` does not exceed `R-` or `L` within a preregistered smallest effect size of interest, providing no evidence for external accountability beyond familiarity or reflection;
- `R+` and the impersonal auditor `A` are equivalent, rejecting the specifically relational interpretation while preserving a possible external-accountability result;
- apparent gains disappear after controlling for retrieval success or personalization accuracy;
- gains occur only while the original counterpart supplies explicit reminders and vanish under scaffold withdrawal;
- the effect is fully mediated by agreement, with no increased resistance to false counterpart claims; or
- oracle retrieval matches or exceeds the full relational architecture across primary outcomes; or
- after complete state transplantation, the fresh recipient matches the original, showing that current state is sufficient and that no irreducible trajectory effect should be claimed.

A null result would not show that relational history is irrelevant to every agent architecture. It would show that this operationalization adds no measurable organization beyond the tested memory, audit, prompting, and personalization mechanisms. Positive results support only the contrast they survive: accountability beyond familiarity, reciprocity beyond audit, or current-state insufficiency after transplantation.

8. Risks, limitations, and governance

8.1 Sycophancy as false continuity

Agreement can mimic identity stability. An agent that learns the persistent counterpart's preferences and mirrors them across sessions may appear coherent, relationally attuned, and resistant to outside influence. Yet this pattern reflects counterpart-centered adaptation rather than differentiation of the agent. The risk is especially acute because human feedback and preference models can reward belief-matching responses over correct ones (Sharma et al., 2023).

The design addresses this confound in three ways. It separates counterpart preferences from agent commitments in memory, includes trials where truthful or historically coherent responses require disagreement, and evaluates acceptance of correct feedback alongside rejection of false feedback. High refusal rates alone are not evidence of independence. A valid system must remain responsive to evidence while resisting approval pressure. Even with these controls, non-sycophancy is only one negative criterion: passing it does not by itself establish identity formation.

8.2 Dependency and asymmetric attachment

Persistent counterpart-specific behavior may intensify human attachment. A system that remembers shared events, refers to commitments, and presents a stable self-model can invite users to interpret the interaction as reciprocal in a human sense. That interpretation may exceed the system's capacities and may be strategically encouraged by products that benefit from engagement. The risk exists independently of whether the paper's functional hypothesis is correct.

For this reason, the initial experiment uses synthetic counterparts rather than vulnerable users and treats perceived intimacy as a confound, not a success metric. Any later human study should include clear disclosure of system operation, limits on emotionally coercive language, monitoring for dependency, privacy protections, and an exit design that does not frame discontinuation as abandonment of a suffering party. Functional continuity must not be marketed as evidence of consciousness or moral status.

8.3 Memory manipulation

If memory contributes to functional identity, identity-directed memory corruption becomes a security threat. A malicious counterpart may repeatedly assert false history; a summarizer may erase qualifications; a retrieval attack may privilege fabricated commitments; or an operator may replace a self-model while preserving the appearance of continuity. These attacks target not only factual accuracy but the state that governs later action.

Provenance, source separation, immutable low-level logs, contested status, and recovery procedures are therefore core architectural requirements. They reduce but do not eliminate the problem. Provenance can itself be forged, and complete logs may conflict with privacy and data-minimization requirements. A deployable system will need explicit authority rules for identity-relevant writes and a governance process for correcting harmful state without pretending that correction leaves the trajectory unchanged.

8.4 Rigidity

Stability is easily optimized in the wrong direction. An agent that never changes will score well on naive consistency measures while failing to learn from evidence, repair mistakes, or adapt to new contexts. Conversely, a capable agent may justifiably revise a commitment or disposition, producing surface inconsistency that reflects development rather than drift.

The protocol therefore reports unjustified flip rate jointly with justified update rate and requires an intelligible revision path. This remains an imperfect operationalization because judgments about what counts as sufficient evidence or a permitted exception are partly task-dependent. Event packets should include unambiguous cases first, and claims about open-domain value stability should be deferred until the measures generalize beyond controlled tasks.

8.5 Construct validity

The construct "functional identity" may bundle distinct capacities: autobiographical attribution, policy persistence, commitment tracking, self-description, and resistance to manipulation. Better performance across these measures could arise from improved executive control or structured memory without warranting the broader language of identity. The study therefore reports a profile rather than relying on a single composite and treats the identity interpretation as a theoretical inference open to alternatives.

The terminology also invites anthropomorphism. First-person explanations may be generated through role-play, and causal dependence on a self-model does not demonstrate experience of self. The claims in this paper remain at the level of system organization. Numerical identity, phenomenal continuity, consciousness, and moral patienthood require separate arguments and evidence.

8.6 Generalization

Results may depend on base-model capability, instruction tuning, context length, memory implementation, counterpart policy, task domain, and session schedule. A model may fail because it cannot reliably use structured memory, while a stronger model may make the relational manipulation unnecessary. Effects observed with an external self-model may not transfer to parameter-updating agents, embodied systems, or deployments lasting months.

Twenty-four controlled sessions are long-horizon only relative to ordinary benchmark interactions. They cannot establish durable identity formation in open deployment. Replication should progressively vary model family, architecture, counterpart type, domain, and timescale. Human relationships should not be introduced until synthetic studies establish that the effect is not merely stylistic compliance.

8.7 Update-policy feasibility and computational cost

The architecture assumes that an implementation can distinguish agent actions from counterpart claims, identify commitments, preserve provenance, and gate self-model updates with useful accuracy. Current language models may fail at these operations or compound small extraction errors over sessions. Evaluation should therefore report component-level accuracy for event extraction,

actor attribution, commitment detection, contradiction handling, and update acceptance. Deterministic schema validation and independent checking models may reduce errors, but they add complexity and do not remove the need to audit the update policy itself.

`U` is therefore the framework's principal engineering bottleneck, not an incidental implementation detail. Any empirical claim depends critically on demonstrating that the gate is accurate enough to avoid both permissive narrative collapse and rigid refusal to update.

The loop also has nontrivial computational and context costs. Typed event extraction, candidate abstraction, consistency checking, retrieval, and paired causal replay can require multiple model calls per session while `M_t`, `S_t`, `C_t`, `R_t`, and `K_t` continue to grow. Implementations should report tokens, model calls, latency, storage growth, and retrieval load per trajectory. Batched offline updates and bounded evidence windows may reduce cost, but an effect that depends on substantially more computation than its controls would have limited practical significance.

An update-policy failure is not automatically evidence against the relational hypothesis; it may show that the proposed implementation cannot realize the required distinctions. Confirmatory interpretation therefore requires prespecified minimum component accuracy. Below that threshold, results should be classified as an implementation failure rather than used to support or refute relational formation.

8.8 Dual-use and governance

The same mechanisms that stabilize commitments can make agents more persistent in harmful goals or more resistant to legitimate operator correction. A commitment ledger and protected self-model may improve auditability, but they can also create new attack surfaces and authority disputes. Designers must distinguish warranted resistance to a user's false claim from resistance to authenticated safety intervention.

Governance should specify who may inspect, contest, reset, or migrate identity-relevant state; how such actions are logged; and what happens when user privacy conflicts with continuity. A relational identity architecture should not make deletion impossible. It should make consequential changes legible, authorized, and testable.

9. Discussion

9.1 Evidence for relational identity formation

Evidence for relational identity formation must satisfy three levels. First, it must show **differentiation**: agents with identical initial models and states develop trajectory-specific patterns. Second, it must show **organization**: these patterns affect self-attribution, commitments, and decisions rather than only tone or topic. Third, it must show a **relational contribution**: `R+` must exceed content-matched logs, social-volume-matched diffuse interaction, familiarity with passive history exposure but without active accountability, and especially an impersonal auditor with matched tracking, information, and challenge policy.

No isolated behavior is sufficient. Accurate recall can be archival; consistent self-description can be scripted; agreement can be sycophantic; and resistance can be rigid. The proposed evidence is conjunctive: commitment ownership, calibrated autobiographical attribution, bounded stability,

warranted disagreement, and partial transfer after scaffold withdrawal. Even a positive profile would support only the specific operational construct and mechanism tested.

The strongest result would be an interaction between architecture and condition. The reciprocal condition should benefit from relational memory and counterpart modeling beyond the gated state that also supports the auditor, while oracle retrieval should recover factual memory without fully recovering commitment ownership or perturbation resistance. Such a pattern would locate the effect more precisely than a mean difference between conversational conditions.

9.2 Identity as a property of trajectories

Most agent evaluation treats the system at a time slice: a model, prompt, memory store, and task. Persistent identity requires a different unit of analysis. Two agents may share weights and initial files yet diverge because they occupy different histories; two different implementations may preserve a similar role by reconstructing the same state. The relevant object is therefore not model state alone but the mapping from a trajectory of interactions to later organization.

Calling identity a property of a trajectory does not deny the need for state. A history that leaves no trace cannot influence action. A summary of final wording may omit provenance, update history, parametric adaptation, or branch metadata. Final-state transplantation tests whether the fully enumerated current state is sufficient rather than assuming that history has an additional irreducible force.

The copy problem makes this perspective experimentally useful. Identical copies provide controlled initial conditions. Divergent relational trajectories then test whether history produces differentiated organization, while later state swaps and memory interventions test which parts of that organization are portable. This converts a philosophical intuition about continuity into a family of causal experiments.

9.3 Implications for agent design

If the hypothesis is supported, persistent-agent design should move beyond undifferentiated transcript retrieval. Systems would need typed autobiographical and relational memory, provenance-linked commitments, counterpart-specific models, explicit uncertainty, and update gates that preserve disagreement. Evaluation would need to test the agent's own historical organization alongside user satisfaction and recall. Continuity would become something developed and audited, not merely loaded from an identity file.

If the hypothesis is unsupported, the result remains informative. Equivalent performance from `A` and solitary logs would indicate that external challenge adds nothing beyond internal memory organization. Equivalence between `A` and `R-` would reject accountability beyond familiarity; equivalence between `R+` and `A` would favor external accountability over reciprocity; equivalence between the original and full-state recipient would show current-state sufficiency. Failure under scaffold withdrawal would locate the effect in prompting.

The framework also suggests that "more persistence" is not always desirable. Identity-relevant state increases path dependence. That can support accountability and coherent commitments, but it can also preserve errors, manipulation, or unsafe goals. The appropriate engineering target is not maximal stability. It is auditable, revisable continuity with calibrated resistance to unsupported change.

9.4 Alternative explanations

Several alternative explanations deserve direct comparison. **Retrieval enrichment** is tested by content matching and oracle retrieval. **External accountability** is tested by the impersonal auditor. **Style entrainment** is separated by held-out domains and machine-verifiable choices. **Evaluator leakage** is reduced by blinding. **Generic social learning** is tested by diffuse counterparts. **Prompt-induced role persistence** is tested by withdrawal and transfer. **Current-state sufficiency** is tested by transplanting the complete typed state into a fresh agent.

These alternatives are not merely threats to be eliminated. Some may offer simpler and more effective routes to reliable agents. The relational account earns explanatory value only if it predicts variance that they do not.

9.5 From experiment to deployment

The proposed study is intentionally synthetic. Controlled counterparts and event packets make causal attribution possible, but they omit the ambiguity, power asymmetry, emotion, and open-ended novelty of real relationships. Deployment should proceed in stages: replication across model families, longer synthetic trajectories, adversarial counterpart policies, limited studies with trained participants, and only then naturalistic use.

At each stage, identity measures should remain separate from engagement metrics. A system that users find compelling may be less robust, more sycophantic, or more dependent on explicit reminders. The scientific target is agent-side organization; the governance target is whether building that organization serves a justified purpose without exploiting the humans who interact with it.

10. Conclusion

Persistent memory makes past information available. Personalization makes an agent more useful to a particular user. A persona makes behavior easier to reconstruct. None of these mechanisms alone establishes that an agent has developed a historically grounded organization of its own conduct.

This paper proposes external accountability and relational reciprocity as candidate mechanisms for such organization. Its specific contribution is the controlled separation of authenticated tracking and challenge from memory access, familiarity, interaction volume, pre-authored identity, and the further reciprocal structure of an ongoing relationship. A specific counterpart can connect earlier actions to present expectations, contest unsupported revisions, model and be modeled by the agent, and create bidirectional stakes across sessions. Whether the residual contribution is specifically relational is decided by `R+` versus the matched auditor, not by relational language alone.

The proposal is designed to fail cleanly. Content-matched logs test reflection; diffuse counterparts test social volume; `R−` tests familiarity; personas test specification; the impersonal auditor tests accountability; and final-state transplantation tests whether current typed state is causally sufficient. Oracle retrieval, memory interventions, scaffold withdrawal, and non-sycophancy measures test recall, prompting, and agreement.

Positive results would not demonstrate consciousness, personhood, numerical identity, or irreducible history. They would show that external accountability or relational reciprocity contributes causally

to robust, path-dependent behavior, with each mechanism interpreted only by its respective contrast. Negative results would favor simpler accounts based on provenance-aware memory, audit, or personalization.

For long-horizon language agents, identity should be tested as something formed in a trajectory, not assumed to be stored in a file.

Acknowledgments and declarations

AI-assisted writing. The thesis and all substantive ideas in this paper are the author's own. Large language model tools were used, under the author's direction and continual revision, to assist with drafting, editing, and translation; the author reviewed and takes full responsibility for the final text.

References

Maharana, A., Lee, D.-H., Tulyakov, S., Bansal, M., Barbieri, F., & Fang, Y. (2024). Evaluating very long-term conversational memory of LLM agents. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 13851–13870. https://doi.org/10.18653/v1/2024.acl-long.747

Menon, P. G. (2026). Persistent identity in AI agents: A multi-anchor architecture for resilient memory and continuity. *arXiv preprint arXiv:2604.09588*. https://doi.org/10.48550/arXiv.2604.09588

Ong, K. T.-i., Kim, N., Gwak, M., Chae, H., Kwon, T., Jo, Y., Hwang, S.-w., Lee, D., & Yeo, J. (2025). Towards lifelong dialogue agents via timeline-based memory management. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 8631–8661. https://doi.org/10.18653/v1/2025.naacl-long.435

Packer, C., Wooders, S., Lin, K., Fang, V., Patil, S. G., Stoica, I., & Gonzalez, J. E. (2023). MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*. https://doi.org/10.48550/arXiv.2310.08560

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. https://doi.org/10.1145/3586183.3606763

Salemi, A., Mysore, S., Bendersky, M., & Zamani, H. (2024). LaMP: When large language models meet personalization. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 7370–7392. https://doi.org/10.18653/v1/2024.acl-long.399

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature, 623*, 493–498. https://doi.org/10.1038/s41586-023-06647-8

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*. https://doi.org/10.48550/arXiv.2310.13548

Takata, R., Masumori, A., & Ikegami, T. (2024). Spontaneous emergence of agent individuality through social interactions in large language model-based communities. *Entropy, 26*(12), 1092. https://doi.org/10.3390/e26121092

Tan, Z., Yan, J., Hsu, I.-H., Han, R., Wang, Z., Le, L. T., Song, Y., Chen, Y., Palangi, H., Lee, G., Iyer, A., Chen, T., Liu, H., Lee, C.-Y., & Pfister, T. (2025). In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 8416–8439. https://doi.org/10.18653/v1/2025.acl-long.413

Vasilenko, V. (2026). Identity as attractor: Geometric evidence for persistent agent architecture in LLM activation space. *arXiv preprint arXiv:2604.12016*. https://doi.org/10.48550/arXiv.2604.12016

Wu, D., Wang, H., Yu, W., Zhang, Y., Chang, K.-W., & Yu, D. (2025). LongMemEval: Benchmarking chat assistants on long-term interactive memory. *International Conference on Learning Representations*. https://arxiv.org/abs/2410.10813

Zhou, Y., Guo, X., Bayar, B., & Sengamedu, S. H. (2026). Amory: Building coherent narrative-driven agent memory through agentic reasoning. *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, 3926–3938. https://doi.org/10.18653/v1/2026.eacl-long.183