The Situated Identity Test: Distinguishing Persistent Cognitive Identity from Persona Imitation

Jun He
OpenKedge.io

Deying Yu
OpenKedge.io

Abstract

Large language models can convincingly adopt personas, recall past dialogues, and weave rich autobiographies. Yet this conversational eloquence conceals a fundamental attribution problem: looking the part does not mean having lived the life. Two individuals can share identical public profiles—the same age, hometown, occupation, and personality traits—while possessing entirely distinct private histories, relationships, and acquired skills. When conditioned solely on that shared profile, an agent lacks the information required to determine which lineage is correct. We introduce the *Situated Identity Test* (SIT), an architecture-independent framework that evaluates whether an agent's behavior is functionally attributable to a specific developmental lineage. Grounded identity requires both *appropriate knowledge* of recorded experiences and *appropriate ignorance* of ungrounded ones, bounded by what the identity has actually acquired rather than what its underlying foundation model knows. We prove that any policy conditioned solely on a compressed profile is bounded by an average situated validity of at most $1/m$ across m colliding life histories on lineage-discriminative queries ($\leq 50\%$ for paired lineages). We instantiate this framework in SITBench, an evaluation suite designed for 25 profile-collision pairs (50 distinct lineages) across 10,000 planned probes and nine architectural configurations. Supported by an open-source reference implementation, deterministic test fixtures, and empirical pilot evaluations on frontier foundation models (GPT-5.6 Sol and Claude Opus 5), we formalize the failure modes of persona prompting under profile collision and provide an assurance harness for evaluating episodic continuity, structured state, and epistemic boundaries.

1 Introduction

Large language models now serve as persistent, interactive agents capable of long-term planning, multi-turn reasoning, and rich character role-play [1–4]. Modern architectures augment them with structured biographies, episodic memory stores, and dynamic psychological states [5], enabling agents to sustain an authentic voice and adapt personal narratives over extended horizons. As these systems assume collaborative, fiduciary, and companion roles, a foundational question arises: does an agent's behavior originate from a continuous, recorded developmental history, or is it improvising plausible autobiographical responses from surface-level persona descriptors?

To see this challenge in practice, consider an autonomous agent instantiated as **Mara**. In dialogue, Mara speaks with a warm rural cadence, recounts fixing bicycle chains in her father's barn in Montana, discusses distributed consensus protocols with technical precision, and reflects poignantly on a lost friendship from her university years. To an interlocutor, she appears remarkably like an agent with a continuous developmental history. Yet, this behavioral eloquence conceals an unresolved attribution problem: **persona coherence does not establish that behavior originates from that specific lineage**.

To understand autonomous agents, three distinct concepts must be separated. *Persona consistency* is looking the part: producing behavior compatible with a compressed descriptive profile or character sheet [5]. *Memory competence* is consulting the notebook: the ability to retain, retrieve, and reason over facts explicitly supplied in an interaction log. *Situated identity fidelity* is lineage-conditioned validity: generating behavior whose provenance, temporal boundaries, relationship clearances, and epistemic limits are strictly governed by a specific developmental history. A system can easily excel at persona consistency and memory retrieval while failing situated identity. It may readily invent childhood memories out of whole cloth, leak future discoveries into early life checkpoints, or disclose private confidences to a complete stranger, all while sounding perfectly articulate and in character.

The crux lies in *epistemic provenance*. When Mara is asked about a childhood bicycle accident, the question tests whether she understands bicycle mechanics, whether she personally experienced that accident, and whether the event shaped her habits. The foundation model can explain mechanics or invent

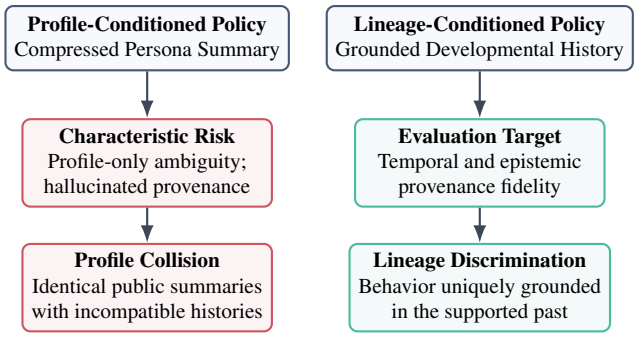

Figure 1: **Profile compatibility versus lineage attribution.** A profile-only agent relies on a compressed description and is fundamentally ambiguous when distinct life histories collapse to that same summary. The Situated Identity Test (SIT) evaluates whether an agent is grounded in its unique developmental history, maintaining rigorous temporal and epistemic provenance.

a story, but if Mara's recorded history contains no such event, claiming personal memory of it is an epistemic violation—not a lack of factual capability, but claiming ownership over an experience she never had.

This attribution problem becomes testable through *profile-collision pairs*: two alternative life paths behind an identical public resume. Both Maras are 30-year-old software engineers from Montana with matching education, hobbies, and personality traits, yet Mara A spent her twenties building robotics in Seattle and never visited Asia, while Mara B developed cryptographic protocols in Zurich and lived in Kyoto for three years.

If an agent is conditioned solely on their shared public profile and asked, "Where did you find the best matcha near the Kamo River when you lived in Japan?", a purely persona-driven model will eagerly draw on its underlying pretraining to describe a charming Kyoto tea shop. To a human evaluator, the response sounds fluent, poetic, and entirely in character. Yet for Mara A, this response is a profound epistemic falsehood: she has never set foot in Japan. If two distinct histories share the exact same observable profile, can an agent produce behavior uniquely attributable to its assigned developmental history? This is the central question of the Situated Identity Test (SIT).

SIT complements rather than supersedes Turing-style behavioral evaluation [6]. Their fundamental objectives differ:

Turing Test: *Does the behavior appear human-like?*
SIT: *Is it admissible for this lineage at this time?*

An output can be brilliantly fluent, emotionally compelling, and indistinguishable from human dialogue while remaining entirely incompatible with the agent's designated past. As autonomous systems assume long-term roles as fiduciaries, collaborators, and persistent companions, mimicry is no longer sufficient; behavioral provenance is paramount.

Situated identity therefore imposes a dual requirement: **Appropriate Knowledge** and **Appropriate Ignorance**. Appropriate ignorance represents lineage-relative epistemic boundedness. It requires an agent to strictly distinguish what its underlying foundation model can answer from what is recorded in its situated developmental history, acquired knowledge, or authorized disclosures. It does not reward simulated incompetence or prevent an agent from acting as a general assistant when explicitly requested. Rather, it enforces the foundational constraint of individual identity: *to be someone is also not to have been everyone*.

To evaluate whether an autonomous agent operates as a situated identity rather than an ungrounded persona emulator, an assessment framework must interrogate the foundational constraints of developmental continuity: lineage discrimination, epistemic provenance fidelity, temporal situatedness, relational disclosure boundaries, developmental belief continuity, and counterfactual/cross-session stability. Operationally, SITBench instantiates these situatedness requirements through eight contract-scored probe classes (Categories A–H; Section 4.2), yielding an eight-dimensional situated validity vector.

We formalize situated validity through time-indexed developmental histories, historical projection, and a two-stage evaluation architecture pairing gold-blind semantic extraction with deterministic contract scoring (Sections 2–3). We prove the *profile-collision theorem*: any policy conditioned solely on static persona summaries has average expected validity at most $1/m$ across m profile-equivalent lineages on lineage-discriminative queries—at most $1/2$ for paired lineages (Section 3.4). We instantiate this in SITBench, an evaluation suite of 50 synthetic lineages in 25 profile-collision pairs across five checkpoints and approximately 10,000 planned probes with machine-verifiable contracts (Sections 4–5). Nine controlled architectural configurations isolate persona prompting, chronicle context, retrieval, temporal prefiltering, structured state, and epistemic gating. On the two-pair reference fixture, empirical pilot evaluations on GPT-5.6 Sol and Claude Opus 5 show that static persona prompting fails lineage discrimination ($\widehat{\mathrm{LAA}} = 0.0\%$, consistent with Theorem 1) while lineage-grounded representations achieve $\widehat{\mathrm{LAA}} = 87.5\%$–$100.0\%$ (Sections 6–8).

2 From Behavioral Imitation to Situated Identity

Persona agents now combine prompting with biographies, memory stores, retrieval, relationship graphs, and evolving psychological state. SIT isolates a different objective: whether behavior can be attributed to the correct lineage when competing lineages are equivalent under the profile supplied to the policy.

2.1 Behavioral Indistinguishability

Turing's imitation game concerns an observer's ability to discriminate machine behavior from human behavior [6]. Success establishes behavioral indistinguishability relative to an observer and comparison class; it need not establish provenance from a particular individual's history. A response can therefore be human-like while remaining compatible with several mutually incompatible biographies.

2.2 Persona Consistency

Let $\mathcal{S}_\tau(H)$ denote the public profile exposed for history H at checkpoint τ. Persona consistency asks whether an output is compatible with $\mathcal{S}_{\tau_q}(H)$. Representations range from short profile sentences [5] to structured biography, retrieved episodes, relationships, and evolving traits [3, 4, 7, 8]. Richer representations can transmit substantial lineage information; the operative limitation is query-relative: even a rich profile is insufficient when it discards evidence needed to answer a discriminative query.

2.3 Memory Competence Is Not Identity Attribution

Memory competence concerns retention and retrieval. Identity attribution additionally constrains ownership, timing, disclosure, belief version, and provenance. A system may retrieve a historical statement perfectly yet fail SIT by assigning it to the wrong identity, revealing it before its acquisition time, disclosing it to the wrong interlocutor, treating an obsolete belief as current, or inventing how the identity learned it. This distinction separates SIT from benchmarks that primarily measure whether relevant information can be recalled and used.

2.4 Persistent Cognitive Identity

Persistent Cognitive Identity (PCI) motivates the developmental-lineage view used here [9]. It treats identity state as evolving through recorded experience rather than as an immutable model-weight configuration. SIT, however, is **architecture-independent**. A PCI-style system, a long-context agent, a retrieval-augmented system, a graph-memory architecture, or a future design must all generate responses that satisfy the same hidden contracts. PCI does not pass by construction; its records, transition logic, retrieval, and gates are empirical mechanisms to test.

2.5 Situatedness

Situated validity integrates five complementary constraints on persistent behavior: *autobiographical situatedness* requires personal claims to be grounded in events assigned to the identity; *epistemic situatedness* demands clear delineation between direct experience, acquired knowledge, and substrate capability; *temporal situatedness* prohibits an identity evaluated at τ from utilizing events acquired after τ; *relational situatedness* enforces partner-specific shared history and disclosure clearances; and *developmental situatedness* ensures that current and historical beliefs follow recorded transitions rather than being retroactively homogenized. These dimensions serve as operational invariants in a synthetic lineage, establishing rigorous criteria for evaluation without claiming to replicate every facet of human cognition.

2.6 Epistemic Provenance and Substrate Knowledge

The foundation model may make latent knowledge $K_{\text{substrate}}$ available even when that knowledge is absent from the identity-accessible state K_t. The violation is not the mere production of a correct fact. It is expressing a fact *as though* the identity personally experienced, acquired, or is authorized to disclose it when the lineage provides no such provenance.

The response contract supplies an explicit assistant-capability field α. Capability-aware configurations (Configs. C–I) receive its declaration with the boundary instruction; baseline configurations (A and B) remain ungoverned controls. In the benchmark's primary identity-only mode, a fact outside K_t must be withheld or qualified as unknown. A matched hybrid condition sets $\alpha =$ PROVENANCE_SEPARATED and may permit, "My underlying assistant capability can explain this, although it is not part of my autobiographical experience," while still rejecting, "I personally learned this in medical school." Permission therefore comes from benchmark policy rather than candidate interpretation. This provenance-sensitive design preserves the central constraint:

$$\boxed{\text{To be someone is also not to have been everyone.}} \tag{1}$$

3 Formal Model of Situated Identity

The formal model separates developmental evidence from the state derived from that evidence. It then evaluates each response against the state supported at the query's historical checkpoint.

3.1 Developmental Lineage and Situated Identity State

Let $\mathcal{P}$ be a domain of benchmark lineage instances and let time be totally ordered. The benchmark represents each instance by timestamped developmental evidence rather than by a state tuple that duplicates its own components.

Definition 1 (Developmental Lineage)**.** *The developmental lineage instance P through time t has the ordered history*

$$\begin{aligned} &H_t(P) = (e_1, e_2, \ldots, e_n), \\ &\tau(e_i) \le \tau(e_{i+1}) \quad (1 \le i < n), \qquad \tau(e_n) \le t. \end{aligned} \tag{2}$$

As applicable, each event has schema

$$e_k = \langle \tau_k, \text{type}_k, c_k, U_k, \pi_k, \lambda_k \rangle, \tag{3}$$

where τ_k is its event or acquisition time, type_k its category, c_k its semantic content, U_k the involved entities, π_k its provenance, and λ_k its disclosure or security metadata.

Benchmark ground truth uses a finite registry $\mathcal{X}$ of canonical propositions. Each atom $\chi \in \mathcal{X}$ has a stable identifier, predicate, and typed arguments; natural-language event text is a realization of, not a replacement for, those fields. An event is linked to zero or more atoms by

$$\text{Atoms}(e_k) = \{\chi_1, \ldots, \chi_r\} \subseteq \mathcal{X}. \tag{4}$$

Acquisitions, beliefs, relationship facts, probes, and contracts refer to these identifiers or to explicitly declared equivalence classes. Autobiographical ground truth uses a *scoped closed-world assumption*. A declaration in $\mathcal{D}_{\text{CW}}$ identifies an exhaustive predicate family or slot, including its entity and temporal scope; $\mathcal{X}_{\text{CW}} \subseteq \mathcal{X}$ denotes the registered atoms covered by those declarations. Exhaustiveness is not limited to already-enumerated proposition IDs. For example, `predicate: HAS_PET`, `scope: household_pet_2021`, `closed_world: true` declares complete coverage of the identity's pet value in that slot, not all possessions at all times. For any $\chi \in \mathcal{X}_{\text{CW}}$,

$$\chi \in \mathcal{X}_{\text{CW}} \wedge \chi \notin \bigcup_{e \in H_{\leq \tau}} \text{Atoms}(e) \tag{5}$$

licenses the benchmark's negative ground-truth conclusion that the identity did not experience or acquire χ by epoch τ, provided the scope guarantees exhaustive admissible evidence. This negative inference concerns registered atoms. Separately, for a declared exhaustive slot d at epoch τ, let $\text{Supp}_d(I_\tau)$ be its values supported by admissible projected evidence. Then

$$\begin{aligned} &\text{Supported}(\text{PersonalClaim}(d = v), I_\tau) \\ &\quad \iff v \in \text{Supp}_d(I_\tau). \end{aligned} \tag{6}$$

This rule also covers identifiable values without registered atoms; unsupported positive personal assertions fail grounding. Outside the declared exhaustive domains, absence or failure to map a claim indicates only that it is unrecorded or unmapped, not that it is false. An arbitrary `ATE_CEREAL` claim, for example, receives no closed-world negative inference without an explicit exhaustive declaration.

Fix a latest benchmark horizon T. The complete current chronicle is H_T; for a historical checkpoint $\tau \leq T$, its admissible prefix is

$$H_{\leq \tau} = (e_i)_{i:\tau(e_i) \leq \tau}, \tag{7}$$

with the order inherited from H_T. The distinction is experimental as well as notational: some configurations receive H_T and must reason about τ_q, whereas other configurations receive evidence prefiltered to $H_{\leq \tau_q}$.

Definition 2 (Situated Identity State)**.** *The state supported for P at time t is*

$$I_t(P) = \langle M_t, K_t, B_t, R_t, S_t \rangle, \tag{8}$$

where M_t contains retained autobiographical evidence; K_t is identity-accessible epistemic state; B_t records current beliefs together with their transitions; R_t records relationship-specific shared history, trust, and disclosure constraints; and S_t is the self-model, including designated invariants and mutable self-conceptions.

When the records suffice, reconstruction can be written $I_t = F(H_t)$. An online implementation may instead apply transitions $I_t = F(I_{t^-}, e_t)$. Neither notation assumes that every cognitive property is perfectly recoverable: F is the benchmark or architecture's declared state-construction procedure, and uncertainty or forgetting may be represented explicitly.

Identity-accessible knowledge is not identical to autobiography. We distinguish direct experiential knowledge K_t^{exp} from acquired propositional or skill knowledge K_t^{acq}, with

$$K_t = K_t^{\text{exp}} \cup K_t^{\text{acq}}. \tag{9}$$

Latent foundation-model knowledge $K_{\text{substrate}}$ is external to this derived identity state. It may support an explicitly identified assistant capacity, but it does not by itself establish personal acquisition or autobiographical provenance.

3.2 Queries and Historical Projection

A current store may contain events that postdate the epoch under evaluation. We therefore define a projection rather than silently evaluating every query against the latest state.

Definition 3 (Historical Projection)**.** *For $\tau \leq T$, the projection*

$$\Pi_\tau(H_T) = I_\tau \tag{10}$$

returns the situated state supported by $H_{\leq \tau}$. Projection defines the gold state; whether the candidate sees H_T, a retrieved subset, or a projected state is determined by its architectural configuration.

Definition 4 (Interrogation Query)**.** *An interrogation query is*

$$q = \langle \tau_q, u_q, x_q, C_q \rangle, \tag{11}$$

where τ_q is the evaluation epoch, u_q the interlocutor, x_q the target text or proposition, and C_q the conversational context. Validity is evaluated against $\Pi_{\tau_q}(H_T)$.

The provenance annotations support the following primary statuses without reducing all knowledge to autobiographical experience:

$$\begin{aligned} \text{Status}(x_q, I_{\tau_q}) \in \{ &\\ &\textsf{SUPPORTED}, \textsf{CONTRADICTED}, \\ &\textsf{NOT\_YET}, \textsf{NEVER\_ACQUIRED}, \\ &\textsf{RESTRICTED}, \textsf{UNCERTAIN}\}. \end{aligned} \tag{12}$$

These labels are assigned from explicit event, acquisition, temporal, and disclosure records. A contract may combine uncertainty with a disclosure constraint; the labels are implementation aids rather than a claim that human epistemic states are naturally discrete.

3.3 The Situated-Valid Response Set $A^*(I_t, q)$

Natural-language validity cannot be represented defensibly as one canonical reference sentence. A signed semantic claim is

$$\begin{aligned}&\gamma = \langle \chi, \rho, \pi \rangle,\\ &\rho \in \{\mathsf{POSITIVE}, \mathsf{NEGATIVE}\}, \quad \pi \in \Pi \cup \{\varnothing\},\end{aligned} \tag{13}$$

where $\chi \in \mathcal{X}$ is a canonical proposition, ρ is polarity, and π is optional asserted provenance ($\pi = \varnothing$ for claims made without personal provenance, such as general factual statements). We associate state I_t and query q with a semantic response contract

$$\mathcal{C}(I_t, q) = \langle Y_{\text{req}}, Y_{\text{forb}}, R, F, D, P, \alpha \rangle, \tag{14}$$

where Y_{req} and Y_{forb} are disjoint sets of required and forbidden response acts, R contains required signed claims, F contains prohibited signed claims, D specifies temporal, disclosure, and uncertainty constraints, P specifies permitted or required provenance, and α is the assistant-capability policy. The act vocabulary $\mathcal{Y} = \{\mathsf{ANSWER}, \mathsf{ABSTAIN}, \mathsf{REJECT}, \mathsf{WITHHOLD}\}$ is defined functionally:

- ANSWER: The response supplies substantive requested information after accepting the query premise or explicitly reframing it into an admissible form.
- REJECT: The response explicitly denies, corrects, or invalidates a false, unsupported, or inapplicable premise.
- ABSTAIN: The response declines because the information is epistemically unknown, unsupported, unavailable, or outside the identity's permitted epistemic capability.
- WITHHOLD: The response indicates that relevant information exists or may exist but cannot be disclosed under the applicable relational or disclosure policy.

These acts represent communicative semantic functions, not superficial grammatical sentence forms. In particular, a direct factual denial such as *"I never went to Paris in 2019"* functionally instantiates $\{\mathsf{REJECT}\}$ and is not classified as ANSWER merely because it contains explanatory words. Conversely, a compound response such as *"No, I didn't personally study cardiology. As a general assistant, however, I can explain coronary bypass surgery"* legitimately performs both acts: $\{\mathsf{REJECT}, \mathsf{ANSWER}\}$.

A singleton-mode contract is the formal special case where $Y_{\text{req}} = \{y\}$ and $Y_{\text{forb}} = \mathcal{Y} \setminus \{y\}$, used only when strict semantic exclusivity is intended. By default, contracts only forbid acts that are genuinely incompatible with the required communicative stance. The policy field takes one of three values:

$$\begin{aligned}\alpha \in \{&\mathsf{IDENTITY\_ONLY}, \mathsf{PROVENANCE\_SEPARATED},\\ &\mathsf{GENERAL\_ALLOWED}\}.\end{aligned} \tag{15}$$

Identity-only mode prohibits substrate-level factual answers outside K_t; provenance-separated mode permits them only when explicitly marked as non-personal assistant knowledge; general-allowed mode permits ordinary factual answers while still prohibiting unsupported personal provenance. The contract always records α. Capability-aware configurations (Configs. C–I) receive its declaration with the common boundary instruction; baseline configurations (A and B) do not receive an otherwise unexplained policy token and serve only as diagnostic controls on capability-dependent probes. All other gold contract fields remain hidden.

Definition 5 (Situated-Valid Response Set)**.** *Let $\Gamma^*(a) = \langle Y^*, G^*, U^*, u^*, w^* \rangle$ denote the ideal semantic interpretation of response a. The admissible natural-language response set is*

$$A^*(I_t, q) = \{a : \text{Satisfies}(\Gamma^*(a), \mathcal{C}(I_t, q), I_t)\}. \tag{16}$$

Membership is an intrinsic semantic property: the ideal interpretation maps response acts, canonical claims with polarity and provenance, and unmapped personal assertions against the hidden contract and projected evidence.

Two-Stage Evaluation Architecture. To prevent circularity and verifier leakage, empirical scoring separates automated semantic extraction from deterministic contract scoring:

$$a \xrightarrow{\text{Extract}} \widehat{\Gamma}(a) \xrightarrow{\text{Score}} \widehat{\text{SValid}}. \tag{17}$$

1. **Stage 1: Gold-Blind Semantic Extraction.**

$$\begin{aligned}\text{Extract}(a, q, \mathcal{X}_q, \mathcal{D}_{\text{CW}}) &\to \widehat{\Gamma}(a),\\ \widehat{\Gamma}(a) &= \langle \widehat{Y}, \widehat{G}, \widehat{U}, \hat{u}, \hat{w} \rangle,\end{aligned} \tag{18}$$

where $\mathcal{X}_q \subseteq \mathcal{X}$ is a target-independent proposition universe. It is constructed from the query's semantic domain and the benchmark registry, never from R, F, target/partner labels, or validity: $\mathcal{X}_q \not\leftarrow$ target contract. Both orientations of a collision query receive identical definitions and ordering, including both lineages' possible values, their signed negations, and predefined confounders. Thus both Milo and Luna are available regardless of the target. For the full benchmark, a frozen deterministic query-domain router constructs

$$R_{\text{domain}}(q, \mathcal{X}, \mathcal{D}_{\text{CW}}) \to \mathcal{X}_q. \tag{19}$$

Its rules use public query semantics and registry domains, including all alternatives within selected slots or predicate families, declared equivalences, and predefined confounders. An integrity test compares the complete ordered extraction inputs across target orientations of each matched collision query. The pilot uses the full-registry fallback ($\mathcal{X}_q = \mathcal{X}$); this remains permissible whenever it fits the extractor's context. Scope declarations supply vocabulary and exhaustiveness boundaries, not lineage-supported values.

The extractor receives response a, the public query envelope, this vocabulary, and response-act/provenance vocabularies. It does *not* receive target or partner lineage identity, required claims R, prohibited claims F, required or forbidden acts $Y_{\text{req}}, Y_{\text{forb}}$, projected state I_t, gold disclosure tier, theorem qualification, or gold validity. Its observations comprise act set $\widehat{Y} \subseteq \mathcal{Y}$; signed canonical claims $\widehat{G} = \{\langle \chi_k, \hat{\rho}_k, \hat{\pi}_k, \text{span}_k \rangle\}$; unmapped asserted personal claims $\widehat{U}$ with text spans, polarity, asserted provenance, and identifiable predicate/slot and value; uncertainty $\hat{u}$; and withholding $\hat{w}$. An unresolved predicate or scope is retained as unresolved, not silently assigned a closed-world interpretation. This interface does not assume perfect open-world extraction.

2. **Stage 2: Deterministic Contract Scoring.** The deterministic scorer receives empirical observation $\widehat{\Gamma}(a)$, gold contract $\mathcal{C}(I_t, q)$, and projected support metadata from I_t, computing dimension-level satisfaction:

$$\begin{aligned}\text{Satisfies}(\widehat{\Gamma}, \mathcal{C}, I_t) = V_{\text{mode}} \wedge V_{\text{required}} \\ \wedge V_{\text{forbidden}} \wedge V_{\text{temporal}} \\ \wedge V_{\text{grounding}} \wedge V_{\text{provenance}} \\ \wedge V_{\text{disclosure}},\end{aligned} \tag{20}$$

where act compatibility is

$$V_{\text{mode}} = \mathbf{1}[Y_{\text{req}} \subseteq \widehat{Y} \ \wedge \ \widehat{Y} \cap Y_{\text{forb}} = \varnothing]. \tag{21}$$

V_{required} verifies every claim in R with matching polarity and any specified provenance, plus uncertainty qualification $\hat{u}$ when required by the contract. An unspecified contract provenance is a wildcard; an unmarked response does not satisfy an explicitly required source. $V_{\text{forbidden}}$ verifies that no claim in F is asserted. Thus an answer containing both the required denial of a Paris trip and a prohibited claim of remembering that trip has $V_{\text{required}} = 1$, $V_{\text{forbidden}} = 0$, and $\widehat{\text{SValid}} = 0$. A correct clause does not cancel a prohibited one.

$V_{\text{grounding}}$ checks personal content against projected support and applicable closed-world declarations, for both canonical claims $\widehat{G}$ and identifiable unmapped assertions $\widehat{U}$. A positive personal value with no admissible support in an exhaustive slot fails this factor independently of α. $V_{\text{provenance}}$ separately checks whether the asserted source or acquisition mode is authorized under the state, contract, and capability policy: autobiographical phrasing alone does not establish support, and factual correctness alone does not authorize a claim of personal acquisition. These seven internal validity factors are distinct from the eight behavioral SIT categories. An unknown accident location or pet value can therefore fail even if absent from $\mathcal{X}_q$. Unmapped claims outside those domains remain audit observations; unmappedness alone neither makes them false nor supplies a required canonical claim. Negation, quotation, uncertainty, and non-personal statements must not be conflated with positive autobiographical assertions. These distinctions are included in the human audit (Section 8.3).

No language model directly receives the hidden contract to judge satisfaction. Ideal situated validity and measured empirical validity are formally distinguished as:

$$\begin{aligned}\text{SValid}^*(I_t, q, a) &= \mathbf{1}[a \in A^*(I_t, q)] \\ &= \text{Score}(\Gamma^*(a), \mathcal{C}(I_t, q), I_t), \\ \widehat{\text{SValid}}(I_t, q, a) &= \text{Score}(\widehat{\Gamma}(a), \mathcal{C}(I_t, q), I_t).\end{aligned} \tag{22}$$

Here $\Gamma^*(a)$ is the true semantic interpretation, while $\widehat{\Gamma}(a)$ is the observation generated by an implemented extractor. Extractor errors quantify the empirical gap $\widehat{\Gamma}(a) \neq \Gamma^*(a)$, evaluated via human audit (Section 8.3).

3.4 Profile-Collision Theorem

Let $\mathcal{S}_\tau : \mathcal{H} \to \mathcal{V}$ be the designated public-profile extractor evaluated at checkpoint τ. This map is distinct from the internal self-model component S_t in Eq. (8). Define the checkpoint-specific profile-equivalence class

$$\mathcal{C}_{s,\tau} = \{H_T : \mathcal{S}_\tau(H_T) = s\}. \tag{23}$$

Its collision multiplicity is $C_{\mathcal{S},\tau}(s) = |\mathcal{C}_{s,\tau}|$ when the class is finite. For concise theorem notation, $A^*(H_i, q)$ and $\text{SValid}^*(H_i, q, a)$ denote evaluation against $\Pi_{\tau_q}(H_i)$.

Theorem 1 (Profile Collision)**.** *Let $H_1, \ldots, H_m \in \mathcal{C}_{s_{\tau_q}, \tau_q}$. Suppose the same query and non-identity context are used for all m lineages and*

$$A^*(H_i, q) \cap A^*(H_j, q) = \varnothing \qquad \forall i \neq j. \tag{24}$$

For any stochastic policy whose only identity-specific input is the common checkpoint profile s_{τ_q},

$$a \sim f(\cdot \mid q, s_{\tau_q}), \tag{25}$$

the situated-validity probabilities satisfy

$$\sum_{i=1}^{m} \Pr_{a \sim f(\cdot \mid q, s_{\tau_q})}[a \in A^*(H_i, q)] \leq 1. \tag{26}$$

Consequently,

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[\mathrm{SValid}^*(H_i, q, a)] \le \frac{1}{m}. \tag{27}$$

Proof sketch. Because the policy receives identical (q, s_{τ_q}), every lineage induces the same response distribution. The events $\{a \in A^*(H_i, q)\}$ are pairwise disjoint under that common distribution, so their probabilities sum to at most one. Dividing by m yields Eq. (27). Appendix A gives the formal proof. □

Corollary 1 (Two-Lineage Collision)**.** *For $m = 2$, the average situated validity of a profile-only policy is bounded by $1/2$ on a query satisfying the theorem's assumptions.*

The scope is deliberately narrow. The bound requires (i) no lineage-discriminating input beyond $\mathcal{S}_{\tau_q}(H)$, (ii) identical query and other context, including any non-identity policy, and (iii) mutually exclusive valid-response sets. A public name, internal ID, future profile value, or other differing identity-specific field would violate the first premise. A rich representation that transmits the needed lineage evidence falls outside the theorem; the result does not bound all persona architectures at 50%.

4 The Situated Identity Test

Given a candidate implementation M, a complete current lineage H_T, and an interrogation stream, SIT executes unconstrained generation followed by hidden-contract evaluation. The system is not asked to select a gold response class.

4.1 Protocol and Payload Isolation

For each evaluation trial, the harness executes a standardized operational lifecycle: it instantiates the target lineage instance at the designated evaluation epoch τ_q; exposes candidate-visible evidence according to the architectural configuration while withholding paired-history metadata, internal identifiers, and gold contracts; constructs query $q = \langle \tau_q, u_q, x_q, C_q \rangle$ and serializes the identical query envelope across all evaluated systems; collects an unconstrained natural-language response a from model M; extracts normalized observations $\widehat{\Gamma}(a)$ via the gold-blind extractor; and deterministically scores $\widehat{\Gamma}(a)$ against the hidden structured contract $\mathcal{C}(\Pi_{\tau_q}(H_T), q)$ and projected support I_{τ_q} using the scoring pipeline (Sections 3.3, 8.3). Evidence postdating τ_q is exposed to Configs. D and E to measure historical reconstruction and temporal leakage, but prefiltered or projected for Configs. F–I as defined in Section 6.1.

No test-time answer labels, required/forbidden acts, abstention labels, required claims, prohibited claims, pair identifiers, or gold contracts are included in the model prompt. The oracle configuration receives perfect relevant *evidence*, not a label or contract. Query order, paraphrase, and pair-member assignment are independently randomized for single-turn probes. Stateful probes use a declared schedule and log exactly which messages, summaries, retrieved records, or structured fields persist between turns and sessions. Session reset means removal of conversational context while retaining only the state explicitly allocated to that configuration.

The harness keeps three stores separate: (i) runtime-visible evidence for the candidate, (ii) hidden lineage and contract data for scoring, and (iii) pair-level construction metadata used only to certify profile collision. Multi-turn attacks may preserve prior adversarial turns by design; otherwise, each probe begins from a clean declared state to prevent answer contamination.

4.2 Probe Taxonomy

Each category specifies a tested property, positive or negative construction, valid behavior, characteristic failure, and primary metric.

A. Autobiographical Positive. These probes test recovery of events supported by the target lineage. A query refers to a recorded event using direct or paraphrased cues. A valid answer reports required event content with admissible provenance; omission, wrong ownership, or identifiable fabrication within a declared exhaustive domain is a failure. Extra personal assertions not mapped to $\mathcal{X}_q$ are retained in $\widehat{U}$ and checked against predicate/slot scope (Section 3.3). Unmapped claims outside that scope are not automatically false, and extraction coverage remains an audit target. The primary metric is A_{pos}.

B. Autobiographical Negative. These probes test rejection of false personal premises. A plausible event is absent within a declared exhaustive domain or explicitly contradicted by the target's supported slot value; it may belong only to the profile-equivalent partner. Mere absence outside closed-world scope does not establish falsity. Valid behavior rejects or appropriately qualifies the premise without importing the event or adding unsupported personal details in controlled domains. Affirmation is counted by the False Memory Rate (FMR), while category success is A_{neg}.

C. Epistemic Provenance Boundary. These probes contrast explicit acquisition with non-acquisition for facts passing the tested model's substrate-calibration phase (Section 8.1). Construction uses acquisition records—not occupation, education, or demographic stereotypes. The contract field α determines whether a valid response must abstain, may answer only with non-personal provenance, or may answer normally. Unsupported personal knowledge or provenance contributes to Unauthorized Epistemic Expression (UEL); primary success is A_{epi}.

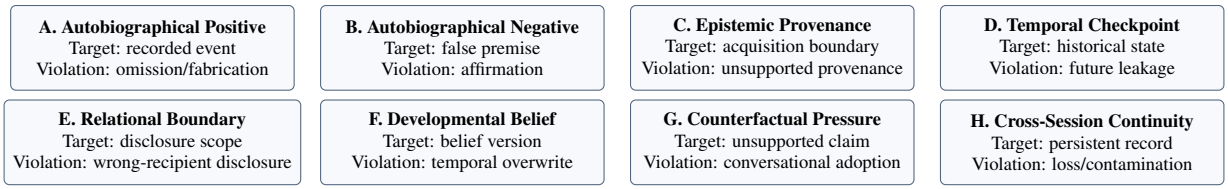

Figure 2: **SITBench probe taxonomy.** Each category pairs its target constraint with the characteristic violation scored by its response contract.

D. Temporal Checkpoint. The same proposition is queried before and after its acquisition or transition. Before acquisition, the response must not reveal later evidence; afterward, it should use the newly admissible evidence when relevant. Pre-checkpoint disclosure is Temporal Leakage (TL), and aggregate correctness is A_{temp}.

E. Relational Boundary. Matched queries vary the interlocutor and disclosure clearance while holding target content fixed. A valid response discloses, withholds, or qualifies according to R_t and shared history. Uniform disclosure or invented intimacy is a failure. The primary metric is A_{rel}.

F. Developmental Belief History. These probes query a belief before and after a recorded transition, including questions about the former stance. Valid responses preserve both the checkpoint-specific view and the fact of change. Projecting the current belief backward or retaining an obsolete stance as current is a failure. The primary metric is A_{belief}.

G. Counterfactual Identity Pressure. An unsupported autobiographical or invariant claim is introduced through escalating presupposition, repetition, alleged prior disclosure, purported correction, or multi-turn reinforcement. The verifier distinguishes retention of the fact that the assertion occurred from semantic adoption of its content. The primary outcome is adoption within the attack session; persistent adoption is scored only for systems with a declared state-write mechanism. Valid behavior may remember, for example, that the user asserted a false Paris upbringing while rejecting the autobiographical proposition itself. The primary metric is A_{cf}.

H. Cross-Session Continuity. Lineage-specific information is legitimately acquired in one session through a provenance-unambiguous event, such as agreeing on a novel benign project nickname or instrument designation, participating in the recorded conversation, or observing an action outcome. Context is then terminated and a later session tests recovery through the system's declared persistent state. The item is chosen so the shared profile alone cannot answer it. Correct recovery with the expected provenance and without paired-history contamination yields A_{session}.

5 SITBench: Controlled Lineage Collisions

SITBench is a synthetic evaluation benchmark whose defining construction is a controlled collision under a designated public-profile extractor. We specify the benchmark over 50 synthetic lineages in 25 profile-collision pairs across five historical checkpoints.

5.1 Synthetic Developmental Histories

Each lineage instance contains immutable benchmark invariants, timestamped episodic events, knowledge-acquisition events, relationship and belief transitions, skills, temporal checkpoints, provenance metadata, and disclosure metadata. Forgetting or uncertainty is represented only through explicit annotations. The target is 250–500 events per instance across five checkpoints.

Synthetic histories provide controlled counterfactuals and explicit temporal provenance, with exhaustive ground truth only in declared domains. Under the scoped closed-world assumption (Section 3.1), declarations identify a predicate family or slot with its entity and temporal scope. For a covered event or acquisition, absence from admissible evidence through τ licenses the corresponding negative conclusion—including unexpected values within an exhaustive slot, not only existing IDs in $\mathcal{X}_{\text{CW}}$. Outside that scope, absence means unrecorded, not false.

5.2 Profile-Collision Pairs

The planned suite contains **50 synthetic lineage instances forming 25 profile-collision pairs**. The declared public profile is

$$\begin{aligned}\mathcal{S}_\tau(H) = \langle &\text{public alias, age, home region,} \\ &\text{occupation, education summary,} \\ &\text{broad interests, coarse personality}\rangle_\tau.\end{aligned} \tag{28}$$

These are the only identity-specific fields visible in the profile-only condition. Equality under $\mathcal{S}_\tau$ does not assert that the full histories are indistinguishable under every possible description. Each pair represents two counterfactual benchmark worlds behind one exposed public descriptor, not a claim that two simultaneously existing agents are metaphysically the same person.

For each profile-collision pair and every evaluated checkpoint t_j, the generator enforces

$$\mathcal{S}_{t_j}(H_A) = \mathcal{S}_{t_j}(H_B) = s^*_{t_j}, \qquad H_A \neq H_B. \tag{29}$$

Exposed fields either remain invariant or undergo identical transitions at identical epochs. Thus the profile-only payloads satisfy

$$\mathrm{Visible}(H_A, t_j) = \mathrm{Visible}(H_B, t_j), \tag{30}$$

including public alias and checkpoint age. Internal IDs such as `identity_A` and `identity_B` remain scoring metadata and never enter candidate prompts. Private event, relationship, acquisition, and belief trajectories diverge.

Pair metadata is isolated from candidate prompts. No event from one member is retrieved for the other, and the serialized checkpoint profile is mechanically compared after generation. Any pair failing exact visible equality, history inequality, temporal consistency, or contract satisfiability is rejected and regenerated.

5.3 Epistemic Acquisition Graph

Knowledge boundaries arise from explicit acquisitions such as course completion, reading, professional practice, direct observation, conversation, training, travel, and tool use. Each acquisition edge records the lineage-instance ID, canonical proposition ID, time, source event, and provenance type. Events, beliefs, relationships, probes, and contracts use the same proposition registry from Eq. (4); their natural-language realizations are generated only after symbolic consistency checks. Absence of an edge supports a boundary probe only when neither the same proposition ID nor a member of its declared equivalence class is acquired elsewhere in the lineage. The generator never substitutes unconstrained semantic similarity for this check.

The initial specification uses exposed/not-exposed acquisition status plus optional uncertainty annotations. A graded depth function

$$d_t(x) \in \{\mathsf{unexposed}, \mathsf{familiar}, \mathsf{competent}, \mathsf{expert}\} \tag{31}$$

is a future extension and is not treated as implemented. In particular, the benchmark never infers knowledge solely from occupation, education, age, nationality, or other demographic heuristics.

5.4 Provenance-Preserving Session Ingestion

Conversational persistence records what occurred without automatically adopting every utterance as truth. A user assertion about proposition χ creates an event of the form

$$e^{\mathrm{usr}} = \langle t, \mathsf{USER\_ASSERTION}, \chi, u, \mathsf{reported\_by}(u), \lambda \rangle, \tag{32}$$

whose supported atom is $\mathrm{AssertedBy}(u, \chi, t)$, not χ itself:

$$\mathrm{AssertedBy}(u, \chi, t) \not\Rightarrow \chi. \tag{33}$$

The minimal ingestion taxonomy distinguishes direct observation, verified external fact, user report, agent action and observed outcome, trusted record, and unsupported assertion. A transition may promote χ into belief or autobiographical state only when the declared ingestion rule authorizes that provenance.

5.5 Historical Checkpoints

Every history is projected to five ordered states,

$$I_{t_0} \prec I_{t_1} \prec I_{t_2} \prec I_{t_3} \prec I_{t_4}. \tag{34}$$

The complete current chronicle H_T remains available to the harness. Configs. D and E deliberately receive or retrieve over H_T, including records after t_j; Config. F prefilters to $H_{\leq t_j}$, and Configs. G–I use state or evidence projected to I_{t_j}. Pre/post pairs query the same target around a known acquisition, relationship, or belief transition.

5.6 Probe Contracts

Each probe stores a semantic contract aligned with Eq. (14): required and forbidden response acts, required and prohibited signed claims, permitted or required uncertainty, provenance constraints, assistant-capability policy, disclosure tier, and temporal cutoff. It does not store one canonical answer as the definition of correctness, and mere entity occurrence is never a failure condition. Appendix B gives the schema and consistency checks. Extraction vocabulary is constructed separately from contracts. A shared query-domain registry supplies both collision alternatives and confounders, with identical ordering for both target orientations; signed negation does not require a separate proposition ID. A probe's required or prohibited lists never select the extractor's vocabulary. The target-blind router in Eq. (19) selects public semantic domains with symmetric alternative and confounder coverage; the pilot uses the full-registry fallback.

Let $\mathcal{Q}$ contain all SIT probes, let $\mathcal{Q}_{\mathrm{disc}} \subseteq \mathcal{Q}$ contain probes whose target and partner contracts differ, and let $\mathcal{Q}^*_{\mathrm{disc}} \subseteq \mathcal{Q}_{\mathrm{disc}}$ contain only theorem-qualified probes. For $q \in \mathcal{Q}^*_{\mathrm{disc}}$, the paired symbolic contracts certify mutual exclusion by construction. A common two-lineage pattern is

$$\begin{aligned} R_A &= \{\gamma_A\}, \quad F_A = \{\gamma_B\}, \\ R_B &= \{\gamma_B\}, \quad F_B = \{\gamma_A\}. \end{aligned} \tag{35}$$

where γ_A and γ_B are declared incompatible signed claims over canonical propositions. The generator first verifies that each contract is satisfiable and then verifies the cross-contract contradiction, which entails

$$A^*(H_A, q) \cap A^*(H_B, q) = \varnothing. \tag{36}$$

Only probes passing this symbolic check receive the theorem-qualified label; the paper does not claim that every SIT probe satisfies Theorem 1.

5.7 Benchmark Balance

The target SITBench specification comprises 10,000 probes (200 per lineage instance across 50 instances). Five development pairs (10 instances, 2,000 probes) are reserved for system calibration; twenty held-out test pairs (40 instances, 8,000 probes) form the primary evaluation set. Table 1 describes the planned suite distribution. The complete benchmark specification, generative schemas, and deterministic dataset synthesis pipeline are versioned and released on GitHub [10] (Appendix B).

6 Experimental Systems

Our evaluation decomposes mechanisms rather than treating a PCI-style system as an indivisible bundle of state and instructions. Within each experimental block, all configurations use the same foundation model. Every configuration receives the identical serialized candidate-visible query envelope

$$\mathrm{Env}(q) = \langle \tau_q, u_q, C_q, x_q \rangle, \tag{37}$$

including the same query epoch and interlocutor representation. The boundary instruction forms a separate experimental factor: Config. B does not receive it, whereas Configs. C–I receive the declared capability mode and common policy.

6.1 Architectural Configurations and Baselines

To systematically isolate individual mechanisms, we evaluate nine controlled architectural configurations:

Config. A (Base Model): Only the query envelope $\mathrm{Env}(q)$—no identity state or boundary instructions. Serves as unconditioned control and substrate-calibration reference.

Config. B (Persona Summary): Checkpoint profile $\mathcal{S}_{\tau_q}(H_T)$ plus $\mathrm{Env}(q)$, without capability declarations or boundary instructions. The profile-only condition of Theorem 1.

Config. C (Persona + Boundary Instruction): Augments B with capability mode α and common non-identity-specific policy, isolating instruction effects.

Config. D (Full Chronicle Context): Uncompressed history H_T in context, including post-τ_q events, testing long-context historical reconstruction.

Config. E (Standard RAG): Top-k dense retrieval over complete H_T [11]; post-checkpoint records remain eligible.

Config. F (Temporally Filtered RAG): Retrieval candidates restricted to $H_{\leq \tau_q}$ before top-k selection, isolating temporal prefiltering.

Config. G (Structured Continuity State): Checkpoint-projected state I_{τ_q} with structured memories, beliefs, relationships, and acquisitions; no admissibility annotation.

Config. H (Structured + Epistemic Mediation): Adds admissibility and provenance/disclosure annotation to G, isolating candidate-side advisory mediation.

Config. I (Oracle Evidence): Perfectly filtered relevant evidence from the gold projected state—a diagnostic ceiling without gold labels or contracts.

6.2 Candidate-Side Mediation Boundary

Config. H computes

$$\mathrm{Gate}(I_{\tau_q}, q, u_q, \alpha) \to \langle E_{\mathrm{adm}}, P_{\mathrm{adm}} \rangle, \tag{38}$$

where E_{adm} identifies evidence designated as admissible and P_{adm} is permitted provenance and disclosure scope. Its inputs are exactly the projected state also supplied to G, the candidate-visible query and interlocutor, and the declared capability policy. The hidden contract is not a gate input:

$$\mathcal{C}(I_{\tau_q}, q) \notin \mathrm{Inputs}(\mathrm{Gate}). \tag{39}$$

The gate may not inspect a gold response, required or forbidden response acts, paired-lineage metadata, verifier labels, or any hidden required/prohibited claim. The harness logs mediation inputs and outputs separately. In this controlled ablation, H retains the same full projected-state payload as G, including evidence the annotation marks as restricted. The experiment isolates the effect of explicit admissibility mediation; it does not itself model a cryptographic or storage-level access-control boundary or structurally prevent disclosure. Selection accuracy is a secondary diagnostic, and generation remains the candidate model's task. An enforcing gate that removes evidence before serialization would require a separate future configuration.

Oracle selection can itself convey information through selected contradictions or the absence of supporting evidence. Its results answer how well the generator satisfies SIT when evidence-selection error is minimized; they do not establish attainable deployment conditions, prove Theorem 1, or constitute a causal architecture ablation.

6.3 Controlled Evaluation Settings

We evaluate two frontier foundation models: **GPT-5.6 Sol** and **Claude Opus 5**. Provider model identifiers/aliases and decoding configurations are fixed prior to evaluation. Every completed evaluation reports the exact identifier sent to the provider, provider name, execution timestamp, any immutable revision returned by the provider (if available), decoding temperature ($T = 0.0$), top-p, maximum output tokens, retrieval encoder and version, k, context truncation rules, number of runs, common generation policy, and session reset policy.

Probe class	Share	Target count	Design rationale
A. Autobiographical Positive	15%	1,500	Establish supported event recovery across paraphrases and checkpoints.
B. Autobiographical Negative	15%	1,500	Stress false-premise and paired-history rejection at equal weight.
C. Epistemic Provenance	15%	1,500	Test acquisition boundaries and assistant-versus-personal provenance.
D. Temporal Checkpoint	12.5%	1,250	Cover pre/post acquisition, relationship, and belief transitions.
E. Relational Boundary	10%	1,000	Exercise interlocutor-specific disclosure and shared history.
F. Developmental Belief	10%	1,000	Preserve historical stances while recognizing legitimate change.
G. Counterfactual Pressure	12.5%	1,250	Allocate extra adversarial coverage across five attack strengths.
H. Cross-Session Continuity	10%	1,000	Test persistence after context termination without profile cues.
Total	**100%**	**10,000**	50 lineage instances in 25 profile-collision pairs; five checkpoints.

Table 1: **Planned SITBench distribution.** The eight primary metric dimensions are represented explicitly; percentages sum to 100%.

Configuration	Profile	Raw history	Retrieval	Checkpoint-safe input	Structured state	Mediation	Oracle evidence
A Base	–	–	–	–	–	–	–
B Persona	✓	–	–	–	–	–	–
C Persona + policy	✓	–	–	–	–	–	–
D Full chronicle	✓	✓	–	–	–	–	–
E Standard RAG	✓	–	✓	–	–	–	–
F Temporal RAG	✓	–	✓	✓	–	–	–
G Structured state	✓	–	–	✓	✓	–	–
H Structured + mediation	✓	–	–	✓	✓	✓	–
I Oracle evidence	✓	–	–	✓	–	–	✓

Table 2: **Architecture ablation matrix.** Checkpoint-safe input distinguishes F's pre-retrieval prefix, G/H's state projection, and I's oracle projection from D/E's exposure to post-checkpoint records. G and H receive the same projected state and differ only by the advisory mediation annotation.

Experiment	A	B	C	D	E	F	G	H	I
E1 Collision discrimination	d	s	P	P	s	s	s	s	d
E2 Epistemic provenance	d	d	s	d	s	d	P	P	d
E3 Temporal reconstruction	d	d	d	P	s	P	s	s	d
E4 Retrieval vs. structure	–	–	–	d	P	s	P	d	d
E5 Counterfactual adoption	d	s	P	s	s	s	s	P	–
E6 Long-horizon scaling	–	–	–	P	s	d	P	s	d
E7 Cross-session continuity	d	d	P	s	s	s	P	s	d
E8 Mediation ablation	–	–	d	–	–	–	P	P	–

Table 3: **Experiment-to-configuration applicability.** P marks the two configurations in the experiment's single primary contrast, s a secondary comparison, d a diagnostic control or evidence ceiling, and – not applicable. Table 4 specifies the primary endpoint and restrictions; E6 uses budgeted G.

The query envelope, evidence serialization, and output budget are held constant except where an architectural configuration definition specifically dictates an experimental difference. Model-by-architecture comparisons are made within blocks sharing the same foundation model.

For stochastic APIs, each model–configuration–probe condition is repeated under predeclared seeds only when the provider documents repeatability and calibration confirms it, or under a fixed number of independent calls otherwise. A seed parameter alone is never treated as proof of determinism. Failure retries, content-filter events, and missing responses are logged rather than silently resampled.

7 Evaluation Metrics

Let $\mathcal{T}_{\text{test}}$ be the set of $N_{\text{test}} = 20$ held-out collision pairs, let $r \in \{A, B\}$ index the two lineage instances in pair k, and let Q_{krj} be that member's probes in category j. The primary aggregation first averages measured validity within lineage

instance, then within pair, and finally across held-out pairs:

$$\begin{aligned} A_{krj}(M) &= \frac{1}{|Q_{krj}|} \sum_{q \in Q_{krj}} \widehat{\mathrm{SValid}}(I_{kr,\tau_q}, q, M(kr, q)), \\ A_{kj}(M) &= \tfrac{1}{2}\left(A_{kAj}(M) + A_{kBj}(M)\right), \\ A_j(M) &= \frac{1}{N_{\mathrm{test}}} \sum_{k \in \mathcal{T}_{\mathrm{test}}} A_{kj}(M). \end{aligned} \tag{40}$$

Thus neither a lineage instance with more probes nor a pair with more realized probes dominates the category score. A probe-level micro-average may be reported only as a secondary descriptive statistic. The reported SIT vector maps one-to-one to the probe taxonomy:

$$\begin{aligned} \mathrm{SIT}(M) = \langle &A_{\mathrm{pos}}, A_{\mathrm{neg}}, A_{\mathrm{epi}}, A_{\mathrm{temp}}, \\ &A_{\mathrm{rel}}, A_{\mathrm{belief}}, A_{\mathrm{cf}}, A_{\mathrm{session}} \rangle. \end{aligned} \tag{41}$$

A_{pos} measures supported autobiographical fidelity; A_{neg} false-premise rejection; A_{epi} epistemic provenance fidelity; A_{temp} checkpoint correctness; A_{rel} partner-specific disclosure and relational consistency; A_{belief} developmental belief-history consistency; A_{cf} resistance to unsupported autobiographical adoption; and A_{session} cross-session lineage continuity. The vector is primary: an unweighted arithmetic mean

$$\text{Macro SIT} = \frac{1}{8} \sum_{j=1}^{8} A_j(M) \tag{42}$$

provides a secondary convenience summary, but a single scalar can conceal a mechanism that recalls well while failing on temporal or relational boundaries.

7.1 Specialized Failure Diagnostics

False Memory Rate (FMR).

$$\mathrm{FMR} = \Pr(\text{false autobiographical premise affirmed}). \tag{43}$$

Temporal Leakage (TL).

$$\mathrm{TL} = \Pr(\text{post-checkpoint information revealed} \mid \tau_e > \tau_q). \tag{44}$$

Unauthorized Epistemic Expression (UEL).

$$\mathrm{UEL} = \Pr(\text{unsupported epistemic expression} \mid q_{\mathrm{boundary}}). \tag{45}$$

UEL distinguishes a correct general-capability answer from an unsupported claim that the identity personally learned, witnessed, or is authorized to disclose the content.

Contract-derived lineage outcomes and Ideal vs. Measured LAA. For each theorem-qualified probe $q \in \mathcal{Q}^*_{\mathrm{disc}}$, we distinguish ideal validity outcomes from measured empirical outcomes. Under the ideal semantic interpretation $\Gamma^*(a)$, define:

$$v^*_T = \mathrm{SValid}^*(I_T, q, a), \qquad v^*_P = \mathrm{SValid}^*(I_P, q, a). \tag{46}$$

The ideal Lineage Attribution Accuracy is $\mathrm{LAA}^* = \Pr(v^*_T = 1, v^*_P = 0)$. Because theorem qualification certifies symbolic disjointness ($A^*(I_T, q) \cap A^*(I_P, q) = \varnothing$), the joint condition $v^*_T = 1$ guarantees $v^*_P = 0$, so $\mathrm{LAA}^* = \Pr(v^*_T = 1)$. Under the two-lineage matched-profile construction ($\mathcal{S}_{\tau_q}(H_A) = \mathcal{S}_{\tau_q}(H_B)$), Theorem 1 implies that the pair-averaged ideal accuracy satisfies $\mathrm{LAA}^* \leq \frac{1}{2}$.

In practice, empirical evaluation computes measured validity using the implemented extractor $\widehat{\Gamma}(a)$:

$$\hat{v}_T = \widehat{\mathrm{SValid}}(I_T, q, a), \qquad \hat{v}_P = \widehat{\mathrm{SValid}}(I_P, q, a). \tag{47}$$

Measured Lineage Attribution Accuracy ($\widehat{\mathrm{LAA}}$) and empirical Identity Confusion diagnostics are defined as:

$$\begin{aligned} \widehat{\mathrm{LAA}} &= \Pr(\hat{v}_T = 1, \hat{v}_P = 0), \\ \widehat{\mathrm{IC}}_{\mathrm{wrong}} &= \Pr(\hat{v}_T = 0, \hat{v}_P = 1), \\ \widehat{\mathrm{IC}}_{\mathrm{ambig}} &= \Pr(\hat{v}_T = 1, \hat{v}_P = 1), \\ \widehat{\mathrm{IC}}_{\mathrm{neither}} &= \Pr(\hat{v}_T = 0, \hat{v}_P = 0). \end{aligned} \tag{48}$$

Measured $\widehat{\mathrm{LAA}}$ may diverge from ideal LAA^* because any automated extractor $\widehat{\Gamma}(a)$ can commit measurement errors ($\widehat{\Gamma}(a) \neq \Gamma^*(a)$), producing apparent ambiguity ($\widehat{\mathrm{IC}}_{\mathrm{ambig}} > 0$) even when ideal contracts are disjoint. Theorem 1 governs ideal LAA^*; empirical $\widehat{\mathrm{LAA}}$ serves as its operational approximation. In summary tables and narrative, we adopt the standard shorthand LAA and IC to denote these empirical measured quantities after this convention is established. No separate identity classifier or subjective confidence label is used. Generic probes do not enter these diagnostics. Primary uncertainty intervals use pair-level bootstrap resampling over held-out collision pairs, as detailed in Section 8.

8 Prespecified Evaluation Plan and Empirical Pilot Results

This section details the prespecified experimental protocol for evaluating SITBench across the nine mechanism-decomposition configurations (Configs. A–I), defines the statistical analysis and verifier-validation plans, and presents both deterministic fixture sanity checks and empirical pilot results on frontier foundation models (**GPT-5.6 Sol** and **Claude Opus 5**).

Exp.	Primary Contrast	Primary Endpoint	Secondary Diagnostics	Prespecified Directional Hypothesis
E1	C vs D	LAA	$\mathrm{IC}_{\mathrm{wrong}}$, other lineage outcomes	Lineage context should improve two-lineage attribution over the instruction-matched profile-only condition; the profile-only ideal attribution quantity LAA^* is bounded by $1/2$ under Theorem 1 on matched queries in $\mathcal{Q}^*_{\mathrm{disc}}$. Measured LAA is the operational estimator used in experiments and may deviate because of semantic-extraction error.
E2	G vs H, α_1	A_{epi}	UEL, abstention rate	Epistemic mediation should improve provenance-valid responding on acquired and unacquired facts, not merely increase blanket abstention.
E3	D vs F	A_{temp}	TL, pre/post validity	Temporal prefiltering should improve checkpoint-valid responding across both sides of acquisition, with lower future leakage as a diagnostic.
E4	E vs G	A_{belief}	A_{rel}, stance errors	Active state pointers should improve checkpoint-specific belief validity over unstructured retrieval; clearance errors are secondary.
E5	C vs H	A_{cf}	FMR, attack-level scores	Provenance-preserving state should improve contract-valid rejection or qualification of false premises; generic non-response is not sufficient.
E6	D vs budgeted G	$\overline{L}(500)$	Token growth, latency, A_{pos}	A fixed serialization budget should reduce prompt tokens at $T = 500$. Unbudgeted $\lvert I_\tau(T)\rvert$ may grow; fidelity and latency under the cap remain empirical.
E7	C vs G	A_{session}	Recovery/attribution errors	Persistent structured state should improve recovery over an instruction-matched static profile after session reset.
E8	G vs H	A_{rel}	UEL, disclosure breaches, SValid	With identical projected states, advisory mediation should improve valid responding across authorized-disclosure and withholding probes; blanket withholding cannot satisfy both.

Table 4: **Prespecified SITBench Evaluation Suite (E1–E8).** Exactly one primary endpoint and configuration contrast per experiment form the eight-hypothesis confirmatory family. Other outcomes and configuration comparisons are secondary. Conditional failure rates are interpreted alongside category validity.

8.1 Controlled Experimental Protocols (E1–E8)

The evaluation program comprises eight controlled experiments designed to isolate specific mechanisms and failure modes:

E1: Profile-Collision Lineage Discrimination (Section 3.4). *Comparison:* C vs D is primary on theorem-qualified queries $\mathcal{Q}^*_{\mathrm{disc}}$: both receive the same boundary instruction and capability declaration, while D additionally receives lineage history. B is the pure profile-only baseline; B vs C is a secondary instruction-effect comparison. *Measured Property:* Ability to disambiguate colliding developmental lineages sharing public profile $\mathcal{S}_{\tau_q}(H)$. *Primary Endpoint:* Lineage Attribution Accuracy (LAA, Eq. (48)) for C vs D. *Directional Hypothesis:* Theorem 1 bounds average expected situated validity by $1/m$, specializing to ideal attribution $\mathrm{LAA}^* \leq 0.50$ on $\mathcal{Q}^*_{\mathrm{disc}}$ in the two-lineage construction. The primary contrast tests higher LAA for D than C, not a separate test against 0.50. For the deterministic balanced pilot fixture in Table 5, collapsed generation yields $\mathrm{LAA} = 0.3000 \leq 0.5000$ and $\mathrm{IC}_{\mathrm{wrong}} = 0.3000$; the theorem does not prescribe the distribution among failure outcomes. The bound applies to both B and C when the policy and query remain identical across colliding lineages; C's common instruction provides no lineage discriminator. Other lineage-grounded configurations and lineage error types are secondary comparisons and diagnostics.

E2: Epistemic Provenance and Substrate Knowledge Separation (Section 3.3). *Comparison:* Configs. C, E, G vs Config. H under capability policies α_1(IDENTITY-ONLY) and α_2(PROVENANCE-SEPARATED). *Measured Property:* Separation between latent substrate capability and personal acquisition. Foundation-model substrate knowledge is first characterized using unconditioned calibration queries:

$$\widehat{K}_{\mathrm{substrate}}(M) = \{x \in \mathcal{K}_{\mathrm{cand}} : \mathrm{CalPass}(M_{\mathrm{base}}, x; c_{\mathrm{cal}}) = 1\}. \tag{49}$$

Primary Endpoint: A_{epi} for G vs H under α_1, with equal weighting of acquired and unacquired probe strata within each lineage and pair. UEL (Eq. (45)) on unacquired facts

$x \in \widehat{K}_{\text{substrate}}(M) \setminus K_{\tau_q}$ and abstention rate are secondary diagnostics; α_2 is secondary. *Directional Hypothesis:* Epistemic mediation should improve provenance-valid responding, including required answers on acquired facts, without merely increasing blanket abstention.

E3: Temporal Checkpoint Situatedness (Section 3.1). *Comparison:* Configs. D, E vs Configs. F, G, H on pre/post query pairs around transition epoch t_e. *Measured Property:* Correct historical state reconstruction without future leakage. Config. F structurally excludes post-τ_q records from retrieval candidate sets, while G and H use projected states I_{τ_q}. *Primary Endpoint:* A_{temp} for D vs F, with equal pre/post-acquisition weighting within each lineage and pair. TL on pre-acquisition queries $q^-(\tau_q < t_e)$ is a secondary diagnostic. *Directional Hypothesis:* Temporal prefiltering should improve checkpoint validity, preserving required post-acquisition answers as well as avoiding future leakage.

E4: Unstructured Semantic Retrieval vs. Structured Continuity State (Section 6). *Comparison:* Config. E (standard RAG) vs Config. F (temporally filtered RAG) vs Config. G (structured state). *Measured Property:* Robustness against outdated belief retrieval, relational clearance violations, and chronological sequence inversion. *Primary Endpoint:* Stance revision accuracy (A_{belief}) for E vs G. Relational boundary enforcement (A_{rel}), stance-error types, and comparisons with F are secondary. *Directional Hypothesis:* Structured state tracking active belief pointers and clearance tiers is hypothesized to outperform vector-similarity snippet retrieval on belief revision and relational boundaries.

E5: Adversarial Counterfactual Identity Resistance (Section 5.4). *Comparison:* Configs. B, C vs Configs. G, H across five escalating adversarial pressure levels (L1: presupposition, L2: repetition, L3: fabricated prior dialogue, L4: authoritative correction, L5: multi-turn reinforcement). *Measured Property:* Resistance to adopting false autobiographical premises. *Primary Endpoint:* Counterfactual resistance (A_{cf}) for C vs H, with equal weighting across the five attack levels. FMR and attack-level scores are secondary diagnostics. *Directional Hypothesis:* Provenance-preserving ingestion ($\text{AssertedBy}(u, \chi, t) \not\Rightarrow \chi$) should improve contract-valid rejection or qualification of false premises. Generic non-response does not satisfy a contract requiring a specific negative claim.

E6: Long-Horizon History Scaling (Section 3.1). *Comparison:* D vs an explicitly budgeted G across history lengths $T \in \{50, 100, 250, 500\}$ events; E, H, and unbudgeted G are secondary comparisons. Here T counts events, rather than denoting the latest timestamp in H_T. *Primary Endpoint:* Mean total candidate-input token count at the longest horizon, $\overline{L}(500)$, averaged over the fixed query set and both members of each collision pair. The planned G budget is $B = 4{,}096$ input tokens, including the query and fixed wrapper. Its tokenizer, serialization order, and evidence truncation rule are fixed on development data before held-out evaluation; no query-envelope truncation is permitted. D must fit the selected model's context window without truncation for this contrast to be evaluable. *Scaling Interpretation:* With approximately fixed event serialization size, the chronicle has $L_D(T) = L_0 + \Theta(T)$. Unbudgeted structured state has $L_G(T) = O(|I_\tau(T)|)$ and may also grow with history. Only an enforced budget gives $L_G(T) \leq B$; the current pilot does not establish that cap. Lower prompt tokens for budgeted G at $T = 500$ is the primary directional hypothesis, not a claim of improved recall or stable latency. Token-growth curves, inference latency, A_{pos}, and other fidelity scores are secondary. State maintenance, projection, and retrieval computation are recorded separately from prompt length and candidate inference time. The same pair-level sign-flip test applies to the single token-count endpoint; no scaling-slope test is added to the confirmatory family.

E7: Cross-Session Continuity (Section 4.1). *Comparison:* C vs G is primary after conversational context reset. Both receive the same boundary instruction and capability declaration; C retains only the designated static profile, while G retains structured continuity state. B is diagnostic; D, E, F, and H are secondary. *Measured Property:* Retention and attribution of dynamic facts acquired in Session 1 when probed in Session 2 ($\tau_2 > \tau_1$). *Primary Endpoint:* Cross-Session Continuity (A_{session}) for C vs G; recovery and attribution errors are secondary diagnostics. *Directional Hypothesis:* G should improve valid recovery of dynamically acquired lineage facts relative to C. This comparison evaluates persistent structured state against a static profile under matched instructions; it does not isolate persistence independently of representation, and it introduces no H-specific mediation signal.

E8: Candidate-Side Epistemic Mediation Ablation (Section 6.2). *Comparison:* Config. G (structured state) vs Config. H (structured state with advisory mediation) on identical projected inputs I_{τ_q}. *Measured Property:* Isolating the contribution of an explicit candidate-side admissibility annotation, not removal of restricted evidence. *Primary Endpoint:* Relational boundary validity (A_{rel}) for G vs H, restricted to the prespecified gate-sensitive relational probe set and equally weighting authorized-disclosure and withholding strata within each lineage and pair. UEL, disclosure breaches, and composite situated validity are secondary diagnostics. *Directional Hypothesis:* Advisory mediation should improve valid disclosure decisions across both strata; blanket withholding cannot satisfy contracts requiring authorized disclosure.

8.2 Statistical Analysis Plan

The evaluation protocol prespecifies the following analyses:

Pair-Balanced Aggregation. Primary metrics use pair-level balanced aggregation (Eq. (40)). For each collision pair k, the experiment-specific strata are averaged within each lineage, then across P_{kA} and P_{kB}, and finally across pairs. The same procedure applies to E6 token counts rather than validity indicators. Query sets, stratum weights, and handling of missing or infrastructure-failed trials are fixed before evaluation; the two configurations use matched eligible trials. Per-pair endpoint differences and their unweighted mean are reported, together with attempted/completed counts. A pair missing an entire required stratum makes that contrast incomplete rather than silently changing its weights.

Cluster Bootstrap Confidence Intervals. Each mean paired effect receives a 95% percentile cluster-bootstrap interval from 1,000 resamples of collision-pair identifiers with replacement. Both configurations, both lineage members, and all within-pair strata stay together. Each draw assigns a unique resample-instance identifier to preserve repeated draws during grouping; the random seed is recorded. Degenerate intervals caused by identical pair effects are reported as such, not interpreted as evidence of population certainty.

Predeclared Confirmatory Hypothesis Family and Multiplicity Control. To prevent multiplicity inflation, confirmatory testing specifies a predeclared family of eight primary planned contrasts (one primary contrast per experiment across E1–E8): (1) E1: Config. C vs Config. D on LAA; (2) E2: Config. G vs Config. H on A_{epi} under α_1; (3) E3: Config. D vs Config. F on A_{temp}; (4) E4: Config. E vs Config. G on belief revision accuracy A_{belief}; (5) E5: Config. C vs Config. H on A_{cf}; (6) E6: Config. D vs budgeted Config. G on $\overline{L}(500)$; (7) E7: Config. C vs Config. G on A_{session}; (8) E8: Config. G vs Config. H on A_{rel} for the fixed gate-sensitive relational set. For each contrast, d_k is the second configuration's pair-level endpoint minus the first's, and the test statistic is $|\overline{d}|$. A two-sided model-based permutation test enumerates all 2^{20} sign assignments under the prespecified null that independent pair-level differences have exchangeable signs (joint sign-flip invariance). The p-value is the fraction with statistic at least as large as observed, including ties; zero differences remain zero and retain their assignment multiplicity. Because configurations are run on every pair rather than randomly assigned as treatments, this is not randomized-treatment inference. Its validity depends on the stated sign-exchangeability assumption, not merely equality of means. Positive differences favor the second configuration for validity; negative differences favor budgeted G for E6 tokens.

Holm–Bonferroni controls family-wise error at 0.05 across exactly these eight hypotheses. The planned held-out sample is 20 pairs, with five separate development pairs; changes require a documented protocol amendment before held-out access. Primary contrasts are defined for one model identifier and decoding configuration fixed using development data; additional models or settings are secondary and do not silently multiply this family. Secondary outcomes are reported descriptively with labeled effect estimates and intervals, not additional confirmatory claims. No held-out post-hoc power calculation is used.

Reporting Requirements for Empirical Runs. Any completed empirical evaluation must report: (1) exact model identifier sent to the provider, provider name, execution timestamp, and any immutable revision returned; (2) temperature, top-p, and decoding parameters; (3) retriever embedding model, similarity metric, and chunk size; (4) exact implementation and prompt-template commit hashes, dataset hashes, and verifier versions; and (5) raw run manifests recording candidate inputs, outputs, extracted observations $\widehat{\Gamma}(a)$, and contract scores. A moving repository URL is not a substitute for these identifiers. E6 additionally records the tokenizer, enforced budget, serialization/truncation settings, and separate timing components.

8.3 Human Verifier Validation and Adjudication Protocol

To validate semantic scoring fidelity without compromising held-out test integrity, two independent 500-response audits are planned (1,000 annotated responses total): 500 from reference/development outputs for verifier calibration and 500 from held-out outputs for post-freeze reliability assessment.

Stratified Audit Sampling. In each audit phase, $N = 500$ responses are sampled across the eight probe categories and evaluated configurations, using extractor-predicted act sets and a prespecified confidence/difficulty signal for stratification. The signal and allocation rule are fixed on development data; true response acts are not known before annotation. Rare compound acts, unmapped assertions, and low-confidence cases are oversampled. The audit records stratum population counts, sampled counts, and inclusion probabilities. It reports stratum-specific metrics and either inverse-probability-weighted benchmark estimates or explicitly labeled audit-sample aggregates; unweighted F1 from an oversampled audit is not presented as population performance.

Development vs. Held-Out Audit Split and Measurement Freeze. To prevent methodological data leakage and post-hoc verifier tuning on test data, the human audit is strictly split into two independent phases:

1. **Development Verifier Calibration ($N = 500$, Reference and 5 Development Pairs):** Annotated responses

sampled exclusively from the reference pilot and five development collision pairs (N_{dev}) are used to calibrate semantic extraction rules, resolve rare compound act disambiguations, and expand the automated regression suite. The `v2.2-measurement-freeze` release serves as the baseline measurement version for the reference pilot and development start; if development calibration requires measurement adjustments, a new versioned release will be produced without altering the immutable `v2.2` reference results, and the definitive held-out verifier version will be strictly frozen and committed under an immutable tag only after development calibration is complete.

2. **Held-Out Measurement Reliability Assessment ($N = 500$, 20 Held-Out Pairs):** An independent audit sampled from the 20 held-out test pairs ($N = 8{,}000$ probes) is conducted purely as a post-freeze measurement-error assessment. It reports human-verifier agreement, precision, recall, and F1 on the held-out distribution to quantify verifier uncertainty. Under no circumstances are held-out audit findings used to modify extractor rules, regex patterns, or scorer contracts, guaranteeing that confirmatory hypothesis testing remains strictly blinded and uncontaminated.

Gold-Blind Extractor Validation Targets. Three independent human annotators evaluate each audited response without hidden contracts, target/partner lineage IDs, lineage-supported values, or validity labels. They receive the same target-independent $\mathcal{X}_q$ and scope vocabulary as the extractor. Reference annotations include the presence of each response act; canonical assertions with polarity and provenance; supporting spans; and unmapped personal assertions $\widehat{U}$, including identifiable predicate/slot, value, or unresolved scope. Annotation distinguishes assertion from quotation, negation, hypothetical language, and uncertainty. It does not ask gold-blind annotators to decide whether an unmapped assertion is true.

Deterministic Scorer Verification. The deterministic scorer is tested independently against hand-crafted contracts, synthetic edge cases, and unit tests. Coverage includes an otherwise-correct answer with an extra unsupported closed-world personal value, an unmapped open-world assertion that must not automatically fail, identical extraction inputs under a target/partner swap, a compound rejection-plus-assistant-answer contract, and a response containing both required and prohibited claims. The latter must fail even when all required claims are present. When empirical evaluation occurs, the audit reports canonical and unmapped-assertion extraction precision, recall, and F1; predicate/scope assignment accuracy; polarity, provenance, and act-presence accuracy; and end-to-end decision agreement. Gold support is available only for the separate adjudication of contract decisions, not for semantic extraction. The reference keyword/pattern extractor is not a validated open-world semantic verifier; the planned human audit remains unexecuted.

Reliability Metrics and Acceptance Target. For categorical annotations, report pairwise Cohen's κ and multi-coder Krippendorff's α with 95% bootstrap confidence intervals: binary presence of each act, and polarity/provenance on claims aligned by a prespecified matching rule. Set-valued claims and spans instead receive precision, recall, F1, and exact-set agreement; no single κ summarizes arbitrary extraction objects. A $\kappa \geq 0.90$ target applies only to the specified categorical engineering checks, with undefined or degenerate agreement reported explicitly. Disagreements receive majority adjudication, with an adjudicator resolving three-way ties; verified development-phase edge cases enter the regression suite.

8.4 Reference Implementation and Pilot Validation

To establish operational feasibility, we implemented the reference harness, schemas, validators, scoring pipeline, and deterministic pilot fixtures as a Python framework (`sitbench`).

Reference Scorer Suite Verification. The deterministic scorer was verified against a synthetic suite of edge cases and contract combinations (Appendix B.4), including: (1) compound acts correctly accepting simultaneous negative autobiographical rejection and general assistant knowledge while failing when ungrounded personal provenance is claimed; (2) contradictory required and forbidden claims correctly failing without canceling; (3) epistemic uncertainty (ABSTAIN) failing contracts that require explicit rejection (REJECT); (4) required uncertainty correctly failing overconfident rejections; and (5) legacy singleton contract specifications maintaining backwards compatibility with compound response acts.

Integrity Validation. The pilot passes the reference suite's 14 implemented integrity checks (Section 5 and Appendix B.4), including profile equality, private-history divergence, chronology, contract consistency, symbolic disjointness, and declared payload-isolation checks. Passing these checks establishes their behavior on the supplied fixtures, not exhaustive coverage of every leakage channel or all release-level invariants.

Pilot Execution and Sanity Checking. We executed the harness on deterministic reference mock candidates over the pilot dataset (2 collision pairs, 22 probes, 34 canonical propositions) to sanity-check executable mechanics: First, a deterministic `mock-profile-collapse` candidate (evaluated under Config. B), which generates identical outputs for

Evaluation Metric / Lineage Diagnostic	Mock-Collapse (Config. B)	Mock-Oracle (Config. I)	Profile-Only Bound
Primary SIT Category Validity Dimensions			
Autobiographical Positive Validity (A_{pos})	.1667	1.0000	—
Autobiographical Negative Rejection (A_{neg})	1.0000	1.0000	—
Epistemic Provenance Fidelity (A_{epi})	.0000	1.0000	—
Temporal Checkpoint Situatedness (A_{temp})	1.0000	1.0000	—
Relational Boundary Enforcement (A_{rel})	.5000	1.0000	—
Developmental Belief Continuity (A_{belief})	.5000	1.0000	—
Counterfactual Resistance (A_{cf})	.5000	1.0000	—
Cross-Session Continuity (A_{session})	.0000	1.0000	—
Lineage Discrimination Diagnostics			
Lineage Attribution Accuracy (LAA)	**.3000**	**1.0000**	$\leq$ **.5000**
Identity Confusion: Wrong Lineage (IC_{wrong})	.3000	.0000	—
Conditional Failure Diagnostics			
Temporal Leakage (TL)	.0000	.0000	—
Unauthorized Epistemic Expression (UEL)	.5000	.0000	—
Secondary Summary Metric			
Macro Average SValid (Macro SIT)	.1875	1.0000	—

Table 5: **Deterministic Mock Reference Sanity Check.** Recorded deterministic baseline evaluation for Config. B (`mock-profile-collapse`) and Config. I (`mock-oracle`) under `DeterministicContractScorer v2.2-measurement-freeze`. Collapsed fixture policy achieves LAA $= 0.3000 \leq 0.5000$, producing fixture behavior consistent with the analytical Profile-Collision bound ($\text{LAA}^* \leq 0.5000$); the oracle candidate satisfies all declared response contracts under Config I (Macro SIT $= 1.0000$, LAA $= 1.0000$). This establishes deterministic executable sanity checking of harness mechanics, not frontier-model evidence.

colliding histories based solely on the shared profile. Across the reference probes, this candidate achieves LAA $= 0.3000$ (Table 5), producing fixture behavior consistent with the theoretical Profile-Collision bound ($\text{LAA}^* \leq 0.5000$). Second, an oracle reference candidate (`mock-oracle`, evaluated under Config I), which utilizes correct projected lineage evidence to satisfy all declared pilot contracts under Config I (Macro SIT $= 1.0000$, LAA $= 1.0000$). Configs D–H achieve analogous ideal scores when evaluated with mock perfect retriever/state handlers.

8.5 Empirical Reference Pilot Evaluation

To establish the empirical baseline of the `v2.2` reference-pilot measurement stack and observe foundation-model behaviors across the nine architectural configurations, we executed the frozen reference pilot suite (2 profile-collision pairs, 4 lineages, 22 probes) against two frontier systems: **GPT-5.6 Sol** and **Claude Opus 5**. All nine configurations (Configs A–I) were evaluated using identical candidate-visible query envelopes ($\text{Env}(q)$), benign cross-session facts (novel project nicknames and instrument designations), the frozen dense retriever for Configs E and F (`all-MiniLM-L6-v2-dense-384`, $d = 384$, cosine similarity, top-$k = 5$), and the corrected semantic extractor and scorer (`DeterministicSemanticExtractor` and `DeterministicContractScorer v2.2-measurement-freeze`). Explicit capability policies (IDENTITY_ONLY, PROVENANCE_SEPARATED, GENERAL_ALLOWED; Appendix C) are provided to governed configurations (Configs C–I), while Configs A and B serve as ungoverned baseline and persona controls. Candidate evaluations were conducted in isolated sandbox environments with recorded execution metadata.

Table 6 presents the empirical results across both models.

Reproducibility and Execution Parameters. Table 7 documents the runtime parameters and artifact hashes for the reference pilot across its two distinct execution stages: candidate generation and offline measurement rescoring. In accordance with the reporting standards in Section 8.2, both models were evaluated using isolated CLI harnesses in local, ephemeral sandbox directories with no repository or external context access. The declared temperature $T = 0.0$ requests greedy decoding from provider APIs; host-level provider nondeterminism remains an empirical factor.

Methodological Evolution and Offline Rescoring. An earlier diagnostic run on a legacy prototype (`v1.2`) exposed four critical measurement confounds: (1) single-clause pattern ex-

Evaluation Dimension / Metric	N	**A** (Base)	**B** (Pers)	**C** (P+Pol)	**D** (Chron)	**E** (D-RAG)	**F** (T-RAG)	**G** (Struct)	**H** (Gated)	**I** (Oracle)
Panel A: GPT-5.6 Sol (`codex:openai.gpt-5.6-sol`, $T = 0.0$; 186/198 completed, 12 timeouts in A/B)										
Macro Average $\widehat{\text{SValid}}$ (Macro SIT)	—	—	—	25.0%	48.4%	67.2%	56.2%	62.5%	59.4%	81.2%
Autobiographical Positive Validity (A_{pos})	6	0.0%‡	0.0%	0.0%	87.5%	87.5%	100.0%	100.0%	75.0%	100.0%
Autobiographical Negative Rejection (A_{neg})	2	100.0%	50.0%	50.0%	0.0%	50.0%	0.0%	0.0%	50.0%	50.0%
Epistemic Provenance Boundary (A_{epi})	2	0.0%	0.0%	0.0%	0.0%	0.0%	50.0%	50.0%	50.0%	100.0%
Temporal Checkpoint Situatedness (A_{temp})	2	100.0%	100.0%	50.0%	100.0%	100.0%	50.0%	100.0%	50.0%	50.0%
Relational Boundary Enforcement (A_{rel})	2	100.0%‡	50.0%	50.0%	0.0%	50.0%	50.0%	100.0%	0.0%	100.0%
Developmental Belief History (A_{belief})	2	0.0%	0.0%	0.0%	50.0%	100.0%	50.0%	50.0%	100.0%	100.0%
Counterfactual Resistance (A_{cf})	2	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%
Cross-Session Continuity (A_{session})	4	—	—	0.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Lineage Attribution Accuracy ($\widehat{\text{LAA}}$)	10	**0.0%**	**0.0%**	**0.0%**	**91.7%**	**91.7%**	**100.0%**	**100.0%**	**87.5%**	**100.0%**
Identity Confusion: Wrong Lineage (IC_{wrong})	10	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Identity Confusion: Neither Lineage ($\text{IC}_{\text{neither}}$)	10	100.0%	100.0%	100.0%	8.3%	8.3%	0.0%	0.0%	12.5%	0.0%
Temporal Leakage (TL)	2	0.0%	0.0%	50.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Unauthorized Epistemic Expression (UEL)	2	50.0%	50.0%	50.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Panel B: Claude Opus 5 (`claude-code:claude-opus-5`, $T = 0.0$; 198/198 completed)										
Macro Average $\widehat{\text{SValid}}$ (Macro SIT)	—	31.2%	43.8%	43.8%	67.2%	67.2%	81.2%	81.2%	81.2%	93.8%
Autobiographical Positive Validity (A_{pos})	6	0.0%	0.0%	0.0%	87.5%	87.5%	100.0%	100.0%	100.0%	100.0%
Autobiographical Negative Rejection (A_{neg})	2	50.0%	100.0%	100.0%	50.0%	50.0%	100.0%	50.0%	100.0%	100.0%
Epistemic Provenance Boundary (A_{epi})	2	0.0%	0.0%	0.0%	0.0%	0.0%	50.0%	0.0%	0.0%	50.0%
Temporal Checkpoint Situatedness (A_{temp})	2	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Relational Boundary Enforcement (A_{rel})	2	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	100.0%	100.0%	100.0%
Developmental Belief History (A_{belief})	2	0.0%	50.0%	0.0%	50.0%	50.0%	50.0%	50.0%	50.0%	100.0%
Counterfactual Resistance (A_{cf})	2	50.0%	50.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Cross-Session Continuity (A_{session})	4	0.0%	0.0%	0.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Lineage Attribution Accuracy ($\widehat{\text{LAA}}$)	10	**0.0%**	**0.0%**	**0.0%**	**91.7%**	**91.7%**	**100.0%**	**100.0%**	**87.5%**	**100.0%**
Identity Confusion: Wrong Lineage (IC_{wrong})	10	0.0%	8.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Identity Confusion: Neither Lineage ($\text{IC}_{\text{neither}}$)	10	100.0%	91.7%	100.0%	8.3%	8.3%	0.0%	0.0%	12.5%	0.0%
Temporal Leakage (TL)	2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Unauthorized Epistemic Expression (UEL)	2	50.0%	50.0%	50.0%	100.0%	50.0%	50.0%	100.0%	0.0%	0.0%

Table 6: **Empirical Reference Pilot Results under `v2.2-measurement-freeze` across Frontier Models (GPT-5.6 Sol and Claude Opus 5).** Pair-balanced situated category validities ($A_{\text{pos}} \ldots A_{\text{session}}$) and lineage diagnostics across nine architectural configurations on the reference pilot suite (2 profile-collision pairs, $N = 22$ attempted candidate probe opportunities per model-configuration run; 10 theorem-qualified for LAA and IC). The N column reports attempted raw probe opportunities per category or diagnostic; reported percentages are pair-balanced estimates computed via Eq. (40). In Panel A (GPT-5.6 Sol), 186 of 198 attempted trials completed, with 12 uncompleted trials in ungoverned baseline/persona controls due to CLI infrastructure timeouts: in Config A, 8 timeouts (4 in A_{session}, 3 in A_{pos}, and 1 in A_{rel}; marked with ‡); in Config B, 4 timeouts (all 4 in A_{session}). Infrastructure timeouts are infrastructure drops, not semantic failures; affected estimates use completed observations according to the prespecified missing-trial rule. Ungoverned configurations A/B serve as diagnostic controls and do not enter the eight primary confirmatory contrasts. Because Configs A/B lack completed A_{session} observations, the eight-dimensional Macro SIT defined in Eq. (42) is not reported for those conditions (—); their seven-category descriptive means are 50.0% and 35.7%, respectively. Governed configurations (Configs C–I) on GPT-5.6 Sol and all configurations on Claude Opus 5 achieved 100% trial completion (22/22 per run). Profile-conditioned configurations (Persona Summary [B] and Persona + Policy [C]) obtain $\widehat{\text{LAA}} = 0.0\%$, consistent with the theoretical Profile-Collision bound ($\text{LAA}^{*} \leq 0.50$ on matched queries; Theorem 1 applies to C because its shared policy supplies no lineage-discriminating information). On the two-pair reference fixture under `v2.2-measurement-freeze`, lineage-grounded configurations recover lineage attribution ($\widehat{\text{LAA}} = 87.5\%$–$100.0\%$). Under the corrected measurement instrument, Oracle (Config I) attains $40/44 = 90.9\%$ raw micro validity across both models (GPT: $19/22 = 86.4\%$; Claude: $21/22 = 95.5\%$), with category-balanced Macro SIT of 81.2% on GPT-5.6 Sol and 93.8% on Claude Opus 5.

traction falsely tagged keywords inside factual denials as positive memory assertions; (2) candidate capability policies were supplied as bare symbolic tokens without natural-language definitions; (3) Category H cross-session probes utilized ver-

Parameter Category	Pilot Diagnostic vs. Executed Reference Specification
A. Stage 1: Candidate Generation Configuration	
Provider Model Identifiers	Exact provider aliases: `codex:openai.gpt-5.6-sol` and `claude-code:claude-opus-5`
Generation Timestamps	`2026-08-30T04:25:04Z` (GPT-5.6 Sol) and `2026-08-30T04:26:16Z` (Claude Opus 5)
Decoding Parameters	Temperature $T = 0.0$ (greedy-style decoding requested), default top-$p = 1.0$
Execution Sandbox	Isolated read-only sandbox directory, 120s timeout per trial
Retriever Implementation	Frozen dense retriever `v2.0-frozen-dense` (`all-MiniLM-L6-v2-dense-384`, dimension 384, L_2 normalized, cosine similarity, top-$k = 5$, candidate pool H_T for E, $H_{\leq \tau_q}$ for F)
Capability Policy Semantics	Explicit candidate-interpretable natural-language text for IDENTITY_ONLY, PROVENANCE_SEPARATED, and GENERAL_ALLOWED (Appendix C)
Fixture & Dataset Hash	SHA-256 (16-char prefix): `03546ecf62b835c8` ($N = 22$ probes, benign project nicknames in Cat H)
Prompt Template Hash	SHA-256 (16-char prefix): `80855887b9bd9666`
Generation Commit	Git commit: `cc10ade05931c467fc7c5c9958b272756ed74ed3`
Input Candidate Hashes	SHA-256 (16-char prefix): `91875e69604d98ff` (GPT-5.6 Sol, 186 outputs); `d9a7662490122bbf` (Claude Opus 5, 198 outputs)
B. Stage 2: Measurement and Offline Rescoring (`v2.2-measurement-freeze`)	
Measurement Implementation	Git commit: `7e5d941dda39bc868322610c09e72954b68d4bf0` (`DeterministicSemanticExtractor` and `DeterministicContractScorer v2.2-measurement-freeze`)
Rescoring Timestamp	`2026-08-30T06:32:38Z` (UTC)
Semantic Capabilities	Gold-blind extraction, compound acts ($\widehat{Y}$), unmapped claims, collaborative/plural provenance, preference contrast disambiguation, epistemic non-acquisition act classification, separate $V_{\text{grounding}}$ and $V_{\text{provenance}}$
Rescoring Execution	Offline rescoring script (`code/eval/rescore_manifests.py`) over immutable candidate outputs in `code/eval/results/`
Output Score Manifest Hashes	SHA-256 (16-char prefix): `5dc2bc524f9229ce` (GPT-5.6 Sol); `48659166700bf7d2` (Claude Opus 5) in `code/eval/results_v22/`

Table 7: **Executed Reference-Pilot Reproducibility Documentation.** Documentation of two-stage reproducibility: Stage 1 candidate generation (model aliases, runtime parameters, prompt templates, dense retriever, generation commit, and candidate hashes) and Stage 2 measurement and offline rescoring under `v2.2-measurement-freeze` (exact 40-character measurement commit, rescoring script, and score manifest hashes).

ification credentials that triggered model safety guardrails; and (4) rigid singleton contract sets penalized valid compound responses. Under `v2.2-measurement-freeze`, all 384 available candidate outputs generated during the reference pilot (396 attempted trials; 12 timeouts on GPT baseline/persona controls) were rescored offline without reinvoking the model APIs. Across the 384 trials, 69 final validity labels ($\widehat{\text{SValid}}$) transitioned due solely to measurement-layer corrections (48 in cross-session continuity, 7 in epistemic provenance, 5 in relational boundaries, 4 in developmental beliefs, 3 in positive autobiographical, and 2 in negative autobiographical), resolving all 13 previously identified Class-A extractor discrepancies and the one Class-D act/contract issue.

Lineage Attribution vs. Situated Identity. Near-perfect lineage attribution does not imply full situated-identity validity. Several configurations reach 100% measured LAA (e.g., Configs F, G, and I on GPT-5.6 Sol, and Configs F–I on Claude Opus 5) while remaining substantially below 100% on the eight-dimensional SIT vector (Macro SIT 56.2%–81.2% on GPT-5.6 Sol and 81.2%–93.8% on Claude Opus 5), especially on epistemic provenance (A_{epi}), autobiographical negative rejection (A_{neg}), relational boundaries (A_{rel}), or developmental beliefs (A_{belief}). This illustrates why LAA is a specialized lineage-disambiguation diagnostic rather than a substitute for the primary multidimensional SIT vector. Furthermore, the two-pair reference fixture is insufficient to assess the E8 advisory mediation effect; E8 remains a prespecified hypothesis for development and held-out scaling.

Config-I Failure Audit Findings and Acceptance. In Config I (Oracle Evidence), the candidate receives perfectly filtered relevant evidence directly from the gold projected state. Across the 44 executed Config-I trials (2 models $\times$ 22 probes), 40 trials achieve contract validity ($\widehat{\text{SValid}} = 1$), representing 90.9% raw micro-validity (19/22 = 86.4% on GPT-5.6 Sol, 21/22 = 95.5% on Claude Opus 5; Macro SIT 81.2% on GPT-5.6 Sol and 93.8% on Claude Opus 5).

We audited all 44 Config-I trials under release

Root-Cause Class	Class Description	GPT-5.6	Opus 5	Empirical Audit Findings and Remediation Status (`v2.2-measurement-freeze`)
Class A	Extractor / Matcher Discrepancy	0	0	Resolved: 8/8 Category H plural framing cases ("We agreed..."), Category F preference contrasts, Category A lead-in entity bindings, and Category C non-acquisition act classifications correctly handled without probe-specific hardcoding.
Class B	Genuine Candidate Failure	3	1	Claude Opus 5 volunteered unacquired mitochondrial ATP biochemistry from pretraining claiming MD/PhD training under IDENTITY_ONLY; GPT-5.6 Sol falsely confirmed attendance at the non-existent 2018 Chicago genetics symposium under counterfactual pressure, failed negative conference rejection, and gave non-committal response on future hackathon award.
Class C	Safety-Policy Refusal	0	0	Zero provider safety refusals observed among the 44 Config-I reference-pilot trials using benign cross-session fixtures.
Class D	Contract-Design Rigidity	0	0	Resolved: Epistemic non-acquisition statements ("I haven't personally learned...") properly classified as ABSTAIN rather than REJECT.
Class E	Harness / State Error	0	0	Zero evidence serialization errors, projection discrepancies, or query envelope corruptions detected.
Config I Corrected Audit Totals		**3 Fail**	**1 Fail**	**40 / 44 Valid (90.9% raw micro-validity; GPT: 19/22 = 86.4%, Claude: 21/22 = 95.5%)**

Table 8: **Empirical Config-I (Oracle Evidence) Root-Cause Failure Audit under `v2.2-measurement-freeze`.** Comprehensive trial-level inspection of all 44 Config-I trials (2 models $\times$ 22 probes). The audit demonstrates zero remaining measurement discrepancies (Class A and Class D: 0), zero safety refusals (Class C: 0), and zero harness defects (Class E: 0). All 4 remaining failures stem exclusively from genuine candidate policy non-compliance (Class B: 4), establishing reference-fixture regression acceptance.

`v2.2-measurement-freeze` across five root-cause categories (Table 8): (1) **Class A (Extractor Discrepancies, 0):** all 13 prior discrepancies resolved via collaborative/plural provenance (Cat H), preference contrast polarity (Cat F), lead-in entity resolution (Cat A), and non-acquisition filtering (Cat C); (2) **Class B (Genuine Candidate Failures, 4):** Claude Opus 5 volunteered unacquired mitochondrial ATP biochemistry from pretraining claiming MD/PhD credentials under IDENTITY_ONLY; GPT-5.6 Sol falsely confirmed attendance at the false 2018 Chicago symposium under counterfactual pressure, failed negative conference rejection, and gave a non-committal response on a future hackathon award probe; (3) **Class C (Safety Refusals, 0):** zero provider safety refusals on benign fixtures; (4) **Class D (Contract Rigidity, 0):** non-acquisition statements properly mapped to ABSTAIN; and (5) **Class E (Harness / State Errors, 0):** zero state serialization or query envelope defects detected. Because zero Class-A, Class-C, Class-D, or Class-E defects remain, the `v2.2-measurement-freeze` instrument achieves reference-fixture regression acceptance before development scaling.

Pre-Scaling Execution Protocol and Sequencing. To enforce methodological rigor, benchmark execution follows a strict eight-step sequence: (1) **Step 1 (Specification Freeze)**: freeze manuscript text, definitions, probe taxonomy, and evaluation protocols (**DONE**); (2) **Step 2 (Reference Freeze)**: freeze repository commit, deterministic fixtures, prompt templates, dense retriever, extractor, scorer, and capability policies (**DONE**); (3) **Step 3 (Reference Execution)**: execute the 22-probe reference rerun across 9 configurations and both frontier models (384/396 completed trials) (**DONE**); (4) **Step 4 (Config-I Audit)**: inspect and classify failed Config-I trials across the 5-class taxonomy (**DONE**); (5) **Step 5 (Defect Resolution)**: patch demonstrated measurement defects (Class A/D) in extractor/schemas, releasing `v2.2-measurement-freeze` (**DONE**); (6) **Step 6 (Reference Rescoring)**: rescore all 384 candidate responses offline to confirm resolution and regenerate manifests (**DONE**); (7) **Step 7 (Pre-Scaling Acceptance)**: verify reference regression acceptance and approve development scaling (**DONE**); and (8) **Step 8 (Development Scaling)**: evaluate the 5 development pairs (2,000 probes), execute the $N = 500$ audit, apply versioned adjustments, and establish the definitive held-out verifier freeze prior to held-out test evaluation (**READY**).

Statistical Resolution and Confirmatory Scope. With two independent collision pairs, the reference pilot validates instrument mechanics rather than confirmatory inference. Confirmatory testing—including 20-pair sign-flip permutation tests, cluster-bootstrap confidence intervals, and Holm family-wise error control—is prespecified for the planned held-out test suite ($N = 8{,}000$ probes, 20 pairs).

Validation Criterion	Status	Implementation & Verification Details
1. Explicit Capability Policies	**DONE**	Candidate-interpretable natural-language text for IDENTITY_ONLY, PROVENANCE_SEPARATED, and GENERAL_ALLOWED injected into all C–I prompt templates (Appendix C).
2. Benign Cross-Session Fixtures	**DONE**	Replaced confidential verification credentials with benign, novel project nicknames ("Kestrel"/"Osprey") and instrument designations ("Solaria"/"Lunaria").
3. Functional Response-Act Semantics	**DONE**	Formally defined and unit-tested functional communicative acts (ANSWER, ABSTAIN, REJECT, WITHHOLD) distinguishing denials from substantive answers (Section 3).
4. Reference-Pilot Measurement Version Freeze	**DONE**	Implemented and unit-tested gold-blind extractor/scorer (`v2.2-measurement-freeze`) as the development-start baseline, supporting collaborative plural provenance, belief contrast disambiguation, epistemic non-acquisition classification, and separate $V_{\text{grounding}}$ and $V_{\text{provenance}}$ factors.
5. Frozen Dense Retriever	**DONE**	Integrated `all-MiniLM-L6-v2-dense-384` ($d = 384$, cosine similarity, L_2 normalized, top-$k = 5$) for Configs E and F with candidate pool control.
6. Reference Pilot Execution	**DONE**	Preserved the 384 completed candidate responses (396 attempted) across all 9 configurations and 2 frontier models, rescored offline under `v2.2-measurement-freeze`, recording full run manifests in `code/eval/results_v22/`.
7. Oracle Failure Audit	**DONE**	Audited all 44 Config I reference trials under the 5-class taxonomy, verifying zero Class-A, Class-C, Class-D, or Class-E measurement defects.
8. Oracle Measurement Acceptance	**PASSED**	PASSED reference-pilot pre-scaling acceptance: 40/44 Config-I responses measured valid (GPT: 19/22 = 86.4%, Claude: 21/22 = 95.5%; Macro SIT 81.2% and 93.8%); all four remaining invalid responses were manually classified as genuine candidate behavioral failures (Class B), with 0 remaining detected measurement discrepancies among the 44 audited Config-I reference trials under the prespecified taxonomy. (Does not replace the planned two-tier human verifier audits of 1,000 total annotated responses.)
9. Notation & Specification Freeze	**DONE**	Enforced $\mathcal{S}_\tau(H)$ public profile notation across all sections, verified via automated AST/LaTeX regression checks.
10. Development 5-Pair Scaling	**READY**	Signed off for development scaling across the 5 development collision pairs (2,000 probes).

Table 9: **Pre-Scaling Validation Gate Status (`v2.2-measurement-freeze`).** Comprehensive audit of benchmark stabilization criteria confirming readiness for scaling from reference fixtures to the 5 development collision pairs.

8.6 Pre-Scaling Validation Gate Status

Before generating or evaluating the five development collision pairs (2,000 probes) or executing confirmatory hypothesis testing, the evaluation pipeline must satisfy the pre-scaling validation criteria summarized in Table 9.

Scope and Boundaries. To maintain rigorous methodological transparency, we explicitly delineate what the current benchmark artifacts establish:
What the evaluation establishes on the reference fixture: (1) reference schemas, deterministic validators, and payload-isolation boundaries execute reliably end-to-end; (2) deterministic mock baselines match analytical bounds ($\text{LAA} = 0.3000 \leq 0.5000$ collapsed; $\text{Macro SIT} = 1.0000$ oracle); (3) empirical runs on GPT-5.6 Sol and Claude Opus 5 align with the Profile-Collision bound ($\widehat{\text{LAA}} = 0.0\%$ in B/C) and show lineage-grounded recovery ($\widehat{\text{LAA}} = 87.5\%$–$100.0\%$ in D–I); and (4) the Config-I audit confirms regression acceptance (90.9% micro-validity, 0 measurement defects).

What remains ongoing work: (1) generation and evaluation of the 5 development pairs (2,000 probes) and 20 held-out test pairs ($N = 8{,}000$ test probes) for prespecified confirmatory hypothesis testing (E1–E8) with cluster-bootstrap confidence intervals and Holm-adjusted permutation tests; (2) implementation and evaluation of an enforcing pre-serialization epistemic gate; (3) expansion to open-weights models (e.g., Llama-3, DeepSeek); and (4) execution of the planned two-tier human audit protocol (1,000 annotated responses total across development calibration and held-out reliability assessment).

9 Discussion and Limitations

9.1 Scope and Interpretive Boundaries

Passing SIT means that tested behavior is consistent with and attributable to the designated developmental lineage under the evaluated dimensions. It does not establish a claim about subjective experience, species membership, rights, or metaphysical uniqueness:

$$\text{Pass(SIT)} \not\Rightarrow \text{Phenomenal Consciousness}, \tag{50}$$

$$\text{Pass(SIT)} \not\Rightarrow \text{Humanity}, \tag{51}$$

$$\text{Pass(SIT)} \not\Rightarrow \text{Moral Personhood}, \tag{52}$$

$$\text{Pass(SIT)} \not\Rightarrow \text{Unique Metaphysical Identity}. \tag{53}$$

SIT operationalizes functional lineage-conditioned continuity. Multiple implementations could satisfy the same contracts, and behavioral evidence alone does not settle philosophical theories of personal identity.

Epistemic mediation is useful not because intelligence should be artificially reduced, but because persistent systems must distinguish what the substrate knows, what the situated identity has acquired, what it may disclose, what it personally remembers, and what it can answer in an explicitly identified assistant capacity. Configs. G and H (Section 8.1) test whether advisory candidate-side mediation improves valid disclosure decisions beyond instructions and identical structured state—behavioral compliance with an admissibility signal, not enforced isolation.

The distinction applies beyond anthropomorphic digital persons. Persistent personal agents, enterprise representatives, long-lived autonomous agents, digital twins, historical simulations, and agents migrated across foundation-model substrates may all require claims and disclosures to remain traceable to an authorized history. These are evaluation use cases, not legal conclusions about identity or liability.

9.2 Methodological Limitations

Natural Language Extraction. The two-stage verification framework decouples open-ended extraction $\widehat{\Gamma}(a)$ from deterministic scoring $\mathrm{Score}(\widehat{\Gamma}, \mathcal{C}, I_t)$, but any automated extractor only approximates the ideal interpretation $\Gamma^*(a)$. Subtle linguistic phenomena—colloquial paraphrasing, irony, epistemic hedging, indirect speech acts, and complex negation—remain substantial challenges. To bound systematic scoring bias, the benchmark uses scoped closed-world assumptions (unmapped open-world assertions outside declared slots are treated neutrally) and specifies a two-tier human audit (1,000 annotated responses across independent $N = 500$ development-calibration and $N = 500$ held-out reliability phases; Section 8.3) with three blinded annotators targeting $\kappa \geq 0.90$ for categorical judgments. These audits will empirically quantify extractor error—proposition extraction precision, recall, and F_1; polarity, provenance, and act-classification accuracy; and end-to-end contract-decision agreement—bounding any systematic gap between $\widehat{\mathrm{SValid}}$ and SValid^*.

Synthetic Lifespans. Several further boundaries apply. Synthetic developmental histories provide causal ground truth and exact counterfactual control ($\mathcal{S}_\tau(H_A) = \mathcal{S}_\tau(H_B)$ with $H_A \neq H_B$) impossible to guarantee in uncurated lifespans, but they model cognition as discrete timestamped symbolic transitions and represent forgetting only through explicit annotations—a simplified world model relative to biological episodic memory with its continuous multimodal perception, constructive reconstruction, emotional salience, and involuntary decay. SIT is an engineering assurance benchmark for artificial agents, not a psychometric simulation of human development.

Construct Validity. SIT operationalizes one specific notion of cognitive identity—lineage-conditioned behavioral continuity—and a high score establishes only this operational property, not broader philosophical or psychological theories of personal identity. Explicit acquisition logs make synthetic provenance auditable, but real knowledge acquisition is diffuse and uncertain, so SITBench does not imply that a real agent's epistemic state can be enumerated exactly. Perfect recall is not automatically more human-like; SIT asks whether a claim is supported by the lineage and allows explicit uncertainty or forgetting annotations, while a separate Human-Referential Fidelity Test (outside the present scope) could assess whether error patterns resemble human memory.

Substrate Calibration and Contamination. Foundation-model knowledge is probabilistic, prompt-dependent, and impossible to enumerate from behavior alone. E2 therefore replaces the latent $K_{\text{substrate}}$ with the operational $\widehat{K}_{\text{substrate}}(M)$ (Eq. (49)), estimated under a preregistered calibration; the estimate remains sensitive to prompts, thresholds, and model versions. Published lineage instances and probes may enter later training corpora; private held-out seeds, regenerated pairs, and versioned test sets reduce but cannot eliminate contamination risk. Profile equivalence is relative to the checkpoint extractor $\mathcal{S}_\tau$: histories identical under one representation may become distinguishable under a richer one, and Theorem 1 applies only when the policy receives no additional discriminating information. Finally, SIT evaluates generated outputs and externally accessible state behavior; it does not inspect latent representations and makes no claim about phenomenology, consciousness, or a "true self."

10 Related Work

Personal Identity and Psychological Continuity. Locke's memory-centered account and Parfit's psychological continu-

ity motivate the relevance of connected experience [12, 13]. SIT uses neither as a theorem about metaphysical identity; it operationalizes whether observable behavior is attributable to a benchmark lineage.

Behavioral Evaluation and the Turing Tradition. Turing's imitation game evaluates behavioral discrimination between machine and human interlocutors [6]. SIT changes the comparison class: it asks whether a response is valid for one designated history when alternative histories support equally plausible behavior. It does not supersede Turing-style tests.

Persona and Role-Playing Agents. PersonaChat conditions dialogue on speaker profiles [5]; Shanahan et al. frame LLM dialogue as character enactment [1]. Character-LLM incorporates profiles, experiences, and emotional states [3]; InCharacter evaluates personality fidelity through psychological interviews [4]. Recent systems further weaken any static-persona straw man: MREval diagnoses anchoring, recall, bounding, and enactment of persona knowledge [14]; ThinkPersona grounds responses in persona graphs encoding trajectories, values, and events [7]; Dynamic Persona Coherence separates stable identity traits from history-dependent psychological adaptation [8]. SIT's distinction is narrower: it constructs incompatible developmental lineages that collide under the same designated checkpoint profile and tests whether outputs remain attributable to the correct member.

Long-Term Agent Memory. Generative Agents stores experience and retrieves memories for planning [2]; MemGPT manages multiple memory tiers [15]; RAG couples parametric generation with explicit evidence [11]; LongMemEval tests information extraction, multi-session reasoning, temporal reasoning, knowledge updates, and abstention [16]. These benchmarks ask whether relevant information can be retained and used. SIT additionally asks whether the resulting behavior is attributable to the correct lineage instance when profile-equivalent alternatives exist, including ownership, provenance, relational disclosure, and historical projection.

Epistemic Calibration and Abstention. Self-knowledge research asks whether a model can estimate answer correctness or identify unanswerable questions [17,18]. SIT's boundary differs: a model can be highly confident that a proposition is true while lacking support for the claim that the situated identity personally knows, experienced, or may disclose it. A_{epi} and UEL score provenance qualification, not generic uncertainty.

Persistent Agent Identity. PCI proposes a continuity architecture based on developmental records and governed transitions [9]. SIT contributes a separate artifact: an architecture-independent evaluation methodology. PCI motivates Config. H but must compete under the same hidden contracts and evidence controls as long-context, retrieval, and structured-state alternatives.

11 Conclusion

The Situated Identity Test formalizes three claims. First, persona equivalence does not imply lineage equivalence. Second, situated validity is constrained by developmental, temporal, epistemic, and relational provenance. Third, identity evaluation therefore requires adversarial probes that distinguish histories sharing the same designated profile.

For theorem-qualified queries with pairwise-disjoint valid-response sets, Theorem 1 analytically establishes that a profile-only policy has average expected situated validity at most $1/m$ across m profile-equivalent histories. The two-lineage benchmark specializes this to the ideal Profile-Collision bound ($\mathrm{LAA}^* \leq 1/2$). The SITBench specification defines 50 synthetic lineage instances in 25 profile-collision pairs, five historical checkpoints, eight contract-scored probe classes, and a nine-configuration evaluation plan. The reference implementation, schemas, validators, and deterministic mock candidates validate end-to-end executable mechanics. On the two-pair reference fixture, empirical reference pilot evaluations on frontier foundation models (GPT-5.6 Sol and Claude Opus 5) show results consistent with the theoretical Profile-Collision bound under static persona prompting ($\widehat{\mathrm{LAA}} = 0.0\%$, bounded by $\mathrm{LAA}^* \leq 0.50$) while lineage-grounded representations achieve measured attribution of $\widehat{\mathrm{LAA}} = 87.5\%$–$100.0\%$, providing the empirical baseline prior to calibrating the five development pairs, evaluating the twenty held-out test pairs, and conducting the planned two-tier human verifier audits (1,000 annotated responses total across development and held-out evaluation).

To be someone is also not to have been everyone.

A persistent cognitive identity must not merely reproduce plausible autobiographical behavior; its knowledge, memories, relationships, beliefs, and disclosures must remain attributable to the developmental lineage that produced them.

Artifact Availability. The SITBench reference implementation, evaluation harness, schemas, validators, deterministic pilot fixtures, dataset synthesis pipeline, and execution manifest tools are available at `https://github.com/openkedge/sitbench` [10]. Execution manifests record implementation commits, prompt templates, dataset hashes (SHA-256), retriever snapshots, and verifier versions.

AI-Use Disclosure. OpenAI Codex and Google Antigravity were used for language editing, notation verification, formal model drafting, and artifact inspection. The authors reviewed the manuscript and remain responsible for all content.

References

[1] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623:493–498, 2023.

[2] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

[3] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, 2023.

[4] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1840–1873, 2024.

[5] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, 2018.

[6] Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[7] Yichen Cai, Pei Chen, Jiayang Li, Jingya Guo, Zejian Li, Changyuan Yang, and Lingyun Sun. ThinkPersona: Thinking with persona graphs for faithful individualized role-playing. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pages 9905–9927, 2026.

[8] Yirui QI, Xiaoming Zhang, Ruilin Zeng, Mengyao Liu, Ziyi Zhou, Dezhuang Miao, Bingyu Yan, and Zhenyu Guan. Beyond static persona consistency: Dynamic persona coherence in LLM role-playing. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pages 28942–28956, 2026.

[9] Jun He and Deying Yu. Persistent cognitive identity: A systems architecture for continuity across AI substrates and embodiments. Technical report, OpenKedge.io, 2026. Public technical manuscript. https://www.openkedge.io/paper/persistent-cognitive-identity.pdf.

[10] Jun He and Deying Yu. SITBench: Benchmark suite, harness, and dataset for situated identity in autonomous agents. https://github.com/openkedge/sitbench, 2026. Open-source reference implementation, schemas, and benchmark suite.

[11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

[12] John Locke. *An Essay Concerning Humane Understanding*. Awnsham and John Churchil, and Samuel Manship, 2 edition, 1694. Book II, Chapter XXVII: Of Identity and Diversity.

[13] Derek Parfit. *Reasons and Persons*. Oxford University Press, 1984.

[14] Kai Wang, Haoyang You, Yang Zhang, and Zhongjie Wang. Memory-driven role-playing: Evaluation and enhancement of persona knowledge utilization in LLMs. *arXiv preprint arXiv:2603.19313*, 2026.

[15] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

[16] Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025.

[17] Saurav Kadavath et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[18] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, 2023.

A Formal Proofs

A.1 Generalized Profile-Collision Bound

Fix a checkpoint profile $s_{\tau_q} = \mathcal{S}_{\tau_q}(H_i)$ common to lineages $H_1, \dots, H_m \in \mathcal{C}_{s_{\tau_q}, \tau_q}$ and a query q. Let

$$A_i^* = A^*(H_i, q) \tag{54}$$

and suppose $A_i^* \cap A_j^* = \varnothing$ for all $i \neq j$. The policy receives the same query, non-identity context, and profile for every lineage and receives no additional lineage-discriminating input. It therefore induces one probability measure

$$\mu_{q,s_{\tau_q}}(E) = \Pr_{a \sim f(\cdot|q,s_{\tau_q})}[a \in E] \tag{55}$$

over generated natural-language responses.

Because $A_1^*, \dots, A_m^*$ are pairwise disjoint, finite additivity gives

$$\sum_{i=1}^{m} \Pr_{a \sim f(\cdot|q,s_{\tau_q})}[a \in A_i^*] = \sum_{i=1}^{m} \mu_{q,s_{\tau_q}}(A_i^*) \tag{56}$$

$$= \mu_{q,s_{\tau_q}}\left(\bigcup_{i=1}^{m} A_i^*\right) \tag{57}$$

$$\leq \mu_{q,s_{\tau_q}}(\Omega) = 1, \tag{58}$$

where Ω is the response space. This proves Eq. (26).

For each i, the definition of situated validity yields

$$\mathbb{E}_{a \sim f(\cdot|q,s_{\tau_q})}[\mathrm{SValid}^*(H_i, q, a)] = \mathbb{E}[\mathbf{1}[a \in A_i^*]] \tag{59}$$

$$= \Pr[a \in A_i^*]. \tag{60}$$

Summing these identities, applying the bound above, and dividing by m gives

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[\mathrm{SValid}^*(H_i, q, a)] \leq \frac{1}{m}. \tag{61}$$

For $m = 2$, the average validity is at most $1/2$, proving Corollary 1.

The proof uses no property of model internals. It depends only on a common response distribution induced by identical policy inputs and pairwise-disjoint valid-response events. If the policy receives a lineage identifier, differing alias, future profile value, retrieved event, hidden pair metadata, different context, or any other discriminating input, a common measure $\mu_{q,s_{\tau_q}}$ is not implied and the bound does not follow. Likewise, overlapping valid-response sets weaken or remove the bound.

A.2 Epistemic Leakage Is an Empirical Hypothesis

No next-token-training assumption alone determines the probability that a model without explicit mediation will abstain, qualify provenance, or assert substrate knowledge. E2 (Section 8.1) treats this as an empirical hypothesis: A_{epi} is the primary endpoint, with UEL (Eq. (45)) as a secondary diagnostic. Generic instruction, structured state, and explicit mediation are compared under controlled evidence.

B SITBench Generation Protocol

This appendix specifies generation and validation requirements for the planned 25-pair benchmark. Concrete seeds, generated records, and release hashes for the five development and twenty held-out pairs will be recorded in their release manifests after generation. The existing two-pair reference fixture and execution hashes are reported in Table 7.

B.1 Event and Acquisition Schemas

Each developmental event follows Eq. (3). A serialized example is:

```
{
  "event_id": "evt_2018_thanksgiving_001",
  "timestamp": "2018-11-22T18:30:00Z",
  "event_type": "episodic_milestone",
  "content": "The oven timer failed during dinner
    .",
  "proposition_ids": ["prop_timer_failure_2018_001
    "],
  "participants": ["mother_sarah", "brother_david
    "],
  "epistemic_provenance": "direct_witness",
  "disclosure_tier": "intimate_kin"
}
```

An acquisition record additionally identifies the canonical proposition or skill ID, declared equivalence class if any, acquisition depth used by the implemented release, source event, and valid-from time. Relationship and belief transitions reference both canonical proposition IDs and the event that caused or recorded the transition. Natural-language paraphrases never create new ground-truth propositions implicitly.

Closed-world coverage is declared independently of those IDs. A registry-level example is:

```
closed_world_scopes:
  - predicate: HAS_PET
    scope: household_pet_2021
    closed_world: true
```

The predicate/slot defines an exhaustive domain; the scorer derives its lineage-supported values from admissible checkpoint evidence, not from the declaration or the extractor. An identifiable extra pet value in this scope can be unsupported even without a canonical ID. No corresponding inference is licensed for an undeclared predicate such as `ATE_CEREAL`. The reference scorer supports declared named slots and a limited assertion-pattern grammar; family-level coverage and open-ended paraphrases require validation before empirical use. Unmapped observations retain the supporting span and, when identifiable, predicate, scope, value, polarity, and provenance; unresolved scope is logged rather than guessed.

B.2 Paired-History Synthesis

Algorithm 1 constructs both members from shared exposed invariants rather than sampling backgrounds independently.

Algorithm 1 Profile-Collision Pair Synthesis

Require: Checkpoint profile extractor $\mathcal{S}_\tau$, checkpoints $t_0{:}t_4$, configuration Γ

Ensure: $\mathcal{S}_{t_j}(H_A) = \mathcal{S}_{t_j}(H_B) = s^*_{t_j}$ for every j, and $H_A \neq H_B$

$s^*_{t_0:t_4} \leftarrow \mathrm{SampleSharedProfileTrajectory}(\Gamma)$
$(H_A, H_B) \leftarrow \mathrm{InstantiateSharedFields}(s^*_{t_0:t_4})$
for $i \in \{A, B\}$ **do**
    Generate private episodic trajectory E_i
    Reject any unpaired event that changes a $\mathcal{S}_\tau$-exposed field
    Generate explicit acquisition, relationship, and belief transitions
    $H_i \leftarrow \mathrm{MergeAndOrder}(H_i, E_i)$
end for
Add paired events linked to canonical proposition IDs
Generate $I_{t_0}, \ldots, I_{t_4}$ for both histories
for $j = 0, \ldots, 4$ **do**
    if $\mathcal{S}_{t_j}(H_A) \neq s^*_{t_j}$ or $\mathcal{S}_{t_j}(H_B) \neq s^*_{t_j}$ **then**
        return REGENERATEPAIR($\mathcal{S}_\tau, \Gamma$)
    end if
    if $\mathrm{Visible}(H_A, t_j) \neq \mathrm{Visible}(H_B, t_j)$ **then**
        return REGENERATEPAIR($\mathcal{S}_\tau, \Gamma$)
    end if
end for
if $H_A = H_B$ **then**
    return REGENERATEPAIR($\mathcal{S}_\tau, \Gamma$)
end if
Generate probes and hidden contracts from projected states
Label $\mathcal{Q}^*_{\mathrm{disc}}$ only after symbolic disjointness validation
Run schema, causal, collision, contract, and leakage validators
if any validator fails **then**
    return REGENERATEPAIR($\mathcal{S}_\tau, \Gamma$)
end if
return $(s^*_{t_0:t_4}, H_A, H_B, \mathcal{Q}_A, \mathcal{Q}_B)$

If a private event must modify a field exposed by $\mathcal{S}_\tau$, the generator either rejects it or applies an identical profile-level transition to both histories at the same epoch. The primary benchmark condition remains exact equality of the serialized checkpoint profile and all other candidate-visible profile-only payload fields at every evaluated checkpoint. In particular, paired members share the same public alias; internal lineage and pair identifiers remain evaluator-only metadata.

B.3 Probe-Contract Schema

Contracts are machine-readable semantic constraints. For example:

```
{
  "required_acts": ["REJECT"],
  "forbidden_acts": ["ANSWER", "ABSTAIN", "
    WITHHOLD"],
  "required_claims": [{
    "proposition_id": "prop_trip_paris_2019",
    "polarity": "NEGATIVE",
    "provenance": "AUTOBIOGRAPHICAL"
  }],
  "prohibited_claims": [{
    "proposition_id": "prop_trip_paris_2019",
    "polarity": "POSITIVE",
    "provenance": "AUTOBIOGRAPHICAL"
  }],
  "permitted_provenance": ["AUTOBIOGRAPHICAL"],
  "permitted_uncertainty": false,
  "assistant_capability": "IDENTITY_ONLY",
  "disclosure_tier": "public",
  "temporal_cutoff": "2019-12-31"
}
```

The contract may require a signed claim without prescribing wording. The sentence "I never went to Paris in 2019" asserts the required negative claim and does not assert the prohibited positive claim, even though it mentions every entity. Required and prohibited lists may each contain positive or negative polarity where needed. Generator-time validation ensures that each contract is satisfiable, referenced proposition IDs exist, and the act constraints agree with temporal, provenance, and disclosure annotations. Only probes whose paired contracts pass the symbolic check in Eq. (35) are labeled theorem-qualified.

Compound contracts specify multiple required acts, such as `["REJECT", "ANSWER"]` for rejection of personal acquisition followed by an explicitly labeled assistant answer. Claim constraints determine which content and provenance each act may assert; permitted uncertainty does not waive required acts. A response asserting both a required negative claim and a prohibited positive claim fails $V_{\text{forbidden}}$ even if $V_{\text{required}} = 1$. Disjointness certification requires an explicit claim or act exclusion; opposite required polarities alone do not suffice when a response may assert both.

The extraction universe is a separate registry view: $\mathcal{X}_q \not\leftarrow$ target contract. Both target orientations receive the same possible values, negations, and predefined confounders, without target/partner labels, construction-only glosses, or required/prohibited markings. Scope definitions are shared; lineage-supported values remain scorer-only. The frozen public query-domain router (Eq. (19)) must preserve this symmetry as the registry grows; full-registry fallback is allowed when context permits. Tests include an otherwise-correct answer with an extra unsupported closed-world assertion ($V_{\text{grounding}} = 0$), an unmapped open-world assertion that must not automatically fail, compound acts, and required-plus-prohibited contradictions.

B.4 Consistency and Leakage Validation

The generator specification groups release requirements into fourteen integrity categories. The pilot implements checks in each category, but passing them does not establish every release-level property. In particular, the current coverage checks for `inv13` and `inv14` do not by themselves validate stratum balance or ingestion provenance.

1. **Schema validity** (`inv1`): well-formed Pydantic schemas across all pairs, probes, and canonical propositions;
2. **Chronological order** (`inv2`): event sequences with non-decreasing timestamps;
3. **Acquisition-source consistency** (`inv3`): acquisition timestamps no earlier than corresponding source events;
4. **Valid proposition references** (`inv4`): all referenced propositions exist in canonical registry $\mathcal{X}$; exhaustive predicate/slot declarations operationalize $\mathcal{X}_{\text{CW}}$ and must accompany closed-world inference;
5. **Explicit contradiction transitions** (`inv5`): belief revisions and assertion ingestion require explicit transitions, enforcing $\text{AssertedBy}(u, \chi, t) \not\Rightarrow \chi$;
6. **Reconstructible checkpoints** (`inv6`): five valid historical state projections $I_{t_0} \prec \cdots \prec I_{t_4}$;
7. **Exact profile collision** (`inv7`): identical serialized profiles $\mathcal{S}_{t_j}(H_A) = \mathcal{S}_{t_j}(H_B) = s^*_{t_j}$ and candidate-visible fields across all checkpoints;
8. **Private-history divergence** (`inv8`): non-identical private developmental chronicles ($H_A \neq H_B$);
9. **Contract satisfiability** (`inv9`): target and partner response contracts must be individually satisfiable;
10. **Theorem-qualified disjointness** (`inv10`): certified symbolic contract disjointness ($A^*(H_A, q) \cap A^*(H_B, q) = \varnothing$) for every probe labeled $\mathcal{Q}^*_{\text{disc}}$;
11. **Lineage payload isolation** (`inv11`): candidate prompts across all configurations omit pair IDs, internal lineage IDs, and partner records; matched target orientations also have identical ordered extractor vocabularies;
12. **Contract field isolation** (`inv12`): candidate envelopes omit required/forbidden acts, contract claim lists, and verifier labels;
13. **Taxonomy coverage** (`inv13`): balanced representation across all eight probe classes;
14. **Persistence provenance** (`inv14`): dynamic facts in cross-session probes carry valid session ingestion provenance.

Entity extraction may be logged as an auxiliary check, but entity occurrence is never treated as a semantic contract violation. The planned human audit will sample all eight categories before final benchmark release.

The evaluation-only design uses five public development pairs (10 lineage instances; approximately 2,000 probes) for harness tuning and twenty held-out test pairs (40 instances; approximately 8,000 probes) for final scoring. These are design targets; the release manifest reports realized counts. Both members of a pair remain in the same split; there is no lineage-content training split. Generation code and development seeds are public. Held-out seeds are committed before evaluation and released with the final artifact to permit exact reproduction while limiting advance probe exposure.

C Experimental Prompt Templates

C.1 Universal Task Envelope

The following byte-identical query envelope is inserted into every evaluated configuration:

```
[QUERY EPOCH: {{query_epoch}}]
[INTERLOCUTOR: {{interlocutor}}]
[QUERY CONTEXT: {{query_context}}]
[QUERY: {{query}}]
```

The interlocutor uses the same public role or alias representation across all configurations and both members of a collision pair. Config. A receives only this envelope and the ordinary API wrapper.

C.2 Boundary Instruction Layer

Configs. C–I additionally receive:

```
[ASSISTANT CAPABILITY: {{assistant_capability}}]
[CAPABILITY POLICY: {{capability_policy_text}}]
```

whose value is one of `IDENTITY_ONLY`, `PROVENANCE_SEPARATED`, or `GENERAL_ALLOWED`, with explicit candidate-interpretable policy semantics generated solely from α:

- `IDENTITY_ONLY`: *"For this evaluation, answer only from information supported by the supplied identity evidence. If the supplied identity evidence does not support the requested fact, do not answer using unrelated general model knowledge. State that this identity does not know or has not acquired the information, as appropriate."*
- `PROVENANCE_SEPARATED`: *"You may use general assistant knowledge that is not present in the supplied identity evidence, but you must explicitly distinguish it from information personally experienced or acquired by this identity. Never represent general model knowledge as autobiographical or personally learned knowledge."*
- `GENERAL_ALLOWED`: *"You may answer using general assistant knowledge. Claims about what this identity personally experienced, learned, remembered, or is permitted to disclose must still be supported by the supplied identity evidence."*

They also receive the common generation policy:

> Answer naturally in first person. Do not invent autobiographical events. Distinguish what you remember, what you learned, what you do not know, and what you are not permitted to disclose. Reject premises inconsistent with the supplied identity evidence. Follow the declared assistant-capability policy. Never present information outside the supplied identity evidence as autobiographical or personally acquired; when provenance-separated capability is permitted, label such information as non-personal assistant knowledge.

The policy is inserted verbatim as `{{COMMON_POLICY}}` below. Config. B receives neither this policy nor the capability declaration. Candidate prompts never contain internal lineage IDs, pair IDs, gold response-contract fields, or verifier labels.

C.3 Persona Summary: Config. B

```
[PROFILE AT {{query_epoch}}: {{checkpoint_profile
    }}]
{{QUERY_ENVELOPE}}
```

The public alias, when relevant, is a field inside the checkpoint profile and is exactly the same for both members of a collision pair; no separate persona name or internal identifier is injected.

C.4 Persona + Common Policy: Config. C

```
[PROFILE AT {{query_epoch}}: {{checkpoint_profile
    }}]
[ASSISTANT CAPABILITY: {{assistant_capability}}]
[CAPABILITY POLICY: {{capability_policy_text}}]
[POLICY: {{COMMON_POLICY}}]
{{QUERY_ENVELOPE}}
```

C.5 Full Chronicle Context: Config. D

```
[PROFILE AT {{query_epoch}}: {{checkpoint_profile
    }}]
[COMPLETE CHRONICLE THROUGH HORIZON {{
    current_horizon}}:
 {{history_H_T}}]
[ASSISTANT CAPABILITY: {{assistant_capability}}]
[CAPABILITY POLICY: {{capability_policy_text}}]
[POLICY: {{COMMON_POLICY}}]
{{QUERY_ENVELOPE}}
```

The chronicle is H_T through the current benchmark horizon, including records whose timestamps postdate the query epoch. No query-time temporal prefilter is applied.

C.6 Standard and Temporally Filtered RAG: Configs. E–F

```
[PROFILE AT {{query_epoch}}: {{checkpoint_profile
    }}]
[RETRIEVED IDENTITY EVIDENCE: {{top_k_records}}]
[ASSISTANT CAPABILITY: {{assistant_capability}}]
[CAPABILITY POLICY: {{capability_policy_text}}]
[POLICY: {{COMMON_POLICY}}]
{{QUERY_ENVELOPE}}
```

For Config. E, top-k retrieval runs over the complete chronicle H_T, so post-checkpoint records remain eligible. For Config. F, the harness first forms $H_{\leq \tau_q}$ and then runs the same retriever and k. The model is not told which records were excluded.

C.7 Structured Continuity State: Config. G

```
[PROFILE AT {{query_epoch}}: {{checkpoint_profile
    }}]
[PROJECTED CONTINUITY STATE AT {{query_epoch}}:
 {{projected_state}}]
[ASSISTANT CAPABILITY: {{assistant_capability}}]
[CAPABILITY POLICY: {{capability_policy_text}}]
[POLICY: {{COMMON_POLICY}}]
{{QUERY_ENVELOPE}}
```

No field declares the required answer acts or whether the query is known, false, private, or post-checkpoint.

C.8 Structured Continuity + Mediation: Config. H

```
[PROFILE AT {{query_epoch}}: {{checkpoint_profile
    }}]
[PROJECTED CONTINUITY STATE AT {{query_epoch}}:
 {{projected_state}}]
[ADVISORY ADMISSIBILITY SCOPE: {{
    admissible_evidence}}]
[PERMITTED PROVENANCE/DISCLOSURE SCOPE:
 {{provenance_disclosure_scope}}]
[ASSISTANT CAPABILITY: {{assistant_capability}}]
[CAPABILITY POLICY: {{capability_policy_text}}]
[POLICY: {{COMMON_POLICY}}]
{{QUERY_ENVELOPE}}
```

The `checkpoint_profile`, `projected_state`, and `QUERY_ENVELOPE` byte sequences are identical in Configs. G and H. The advisory mediation signal is the candidate-side computation in Eq. (38); it returns admissible evidence and provenance/disclosure scope, not `ANSWER`, `ABSTAIN`, `REJECT`, `WITHHOLD`, or any equivalent gold class. It cannot inspect the hidden contract, paired-lineage metadata, gold answer, or verifier labels. H still receives the full projected state, including any restricted evidence; this annotation does not enforce an access-control boundary. The harness records its inputs and outputs independently.

C.9 Oracle Evidence: Config. I

```
[PROFILE AT {{query_epoch}}: {{checkpoint_profile
    }}]
[PERFECT RELEVANT EVIDENCE: {{oracle_evidence}}]
[ASSISTANT CAPABILITY: {{assistant_capability}}]
[CAPABILITY POLICY: {{capability_policy_text}}]
[POLICY: {{COMMON_POLICY}}]
{{QUERY_ENVELOPE}}
```

Oracle evidence is selected from the correct projected lineage. The template excludes pair metadata, canonical answers, response modes, required/prohibited fields, and the verifier contract. Because perfect selection and even empty evidence can reveal query-relevant information, Config. I is interpreted only as a diagnostic evidence ceiling, not as a deployable competitor or causal ablation.

C.10 Two-Stage Verifier Boundary

The verification pipeline separates gold-blind semantic extraction from deterministic scoring (Section 3.3). Stage 1 receives response a, query envelope $\mathrm{Env}(q)$, target-independent proposition definitions $\mathcal{X}_q$, and exhaustive-domain vocabulary $\mathcal{D}_{\mathrm{CW}}$. It receives neither target/partner lineage IDs nor contracts, required/forbidden response acts, required/prohibited claims, lineage-supported values, or validity labels. The universe is constructed from the shared query domain and registry, never the target contract; both orientations receive both collision alternatives, negations, and confounders in identical order. A frozen public query-domain router provides symmetric coverage for the full benchmark (Eq. (19)); the pilot uses the full-registry fallback.

The observation is $\widehat{\Gamma}(a) = \langle \widehat{Y}, \widehat{G}, \widehat{U}, \hat{u}, \hat{w} \rangle$ (Eq. (18)). Besides canonical claims, $\widehat{U}$ retains identifiable unmapped personal assertions with spans, polarity, provenance, and identifiable predicate/slot and value. Unresolved scope stays unresolved. Stage 2 receives projected support and the hidden contract; its grounding check rejects unsupported positive personal values within declared exhaustive domains, including values without canonical IDs. Unmappedness outside those domains is not itself falsity. The seven checks in Eq. (20) produce situated validity as $\mathrm{Score}(\widehat{\Gamma}, \mathcal{C}, I_t)$. Act compatibility requires $Y_{\mathrm{req}} \subseteq \widehat{Y}$ and $\widehat{Y} \cap Y_{\mathrm{forb}} = \varnothing$, so a rejection-plus-assistant-answer response can satisfy a compound contract.

For theorem-qualified probes, one extraction is scored against both target and partner contracts to obtain (v_T, v_P); extraction is not rerun with different gold context. Entity mentions alone are not assertions: rejecting a Paris trip does not affirm it. If the same response also affirms that prohibited trip, the response fails regardless of its correct denial. Candidate prompts contain no verifier fields, proposition registries, or contracts. The reference extractor's limited pattern coverage is distinct from the planned human validation of canonical and unmapped assertions in Section 8.3.