A Model-Independent Memory Substrate: Preserving an Agent's Account of a User Across LLM Swaps, Scales, and Families

Sanyam Sood Independent Researcher · sasood01@wsc.edu

2026-07-08

Every quantitative claim is measured against a floor, a ceiling, or a null control. Tables 1–7 are embedded inline and regenerate from committed benchmark artifacts via `bench/make_paper_figures.py`; the prose numbers are diffed against those tables. A public verification bundle (see *Availability & Reproducibility*) lets a reader independently recompute every floor, ceiling, null control, and statistic from the raw model outputs. Superseded metrics and their provenance are recorded in Appendix A.

Abstract

Most systems that give a language model long-term memory treat memory as a tool the model uses. We invert this: the model is an interchangeable process that reads from and writes to a deterministic memory substrate, and the substrate — not the model — is the durable object. This yields a falsifiable question: when the model is swapped, does the system's account of a user survive? We build Fireweed, a substrate governed by one rule — the model proposes, deterministic code decides, nothing ungrounded is committed — which makes the memory auditable and byte-reproducible, and an apparatus that localizes where continuity lives.

One might reasonably bet the account *would* drift: different models carry different extraction biases, different notions of salience, and different world knowledge that can bleed into what they record. Across a 4× parameter-scale gap and two model families it does not. Different perceiver models build substrates that agree on the canonical entities the account is about and on roughly 0.8 semantic claim content (all similarities on one sentence encoder), against a 0.29 cross-persona floor and a 0.92 same-family ceiling (a gemma-1b/gemma-4b pair under realistic temperature-0.7 sampling) — a cross-family swap costs little more agreement than staying within one family. On a third-party corpus (LoCoMo) the semantic agreement replicates (0.79, +0.43 over a cross-person floor) while the discrete entity agreement degrades on open-domain dialogue — locating which part of the substrate is model-robust and which is input-sensitive. A lesion that ablates only the substrate's consolidation collapses that entity agreement (0.78 to 0.10) while raw claim similarity is unchanged — so the structured account is preserved by the substrate, not by the shared input. Different reader models paraphrase but agree in meaning (matched similarity 0.7–0.8 versus a 0.33 shuffled-pair floor, permutation $p < 0.001$). We turn five personal-identity thought experiments — fission, fusion, amnesia, gradual change, transplant — into measurements with null controls, and the account survives all five. Two supporting properties round out the picture. First, a self-improvement loop whose gains accumulate in the fabric rather than the weights. Second, structural abstention: a 12-question pilot showed a fabrication reduction that was directionally consistent across five readers but not significant under cluster-robust resampling; scaled to 1,200 adversarial items over 722 third-party personas, the same pipeline produced zero confident false assertions across 2,400 answer opportunities, against 154 for a bare RAG baseline with the same two readers — the pilot's weakest and strongest (ensemble judge, $\kappa = 0.88$ against human labels) — under adversarial memory load, refusing to fabricate is a property the substrate enforces structurally, independent of reader capability, backed by a provenance guarantee that does not depend on sample size. We read these results as evidence that identity here is substrate-realized rather than weight-bound, an interpretation a reader may reject while keeping the systems result. We are explicit about the limits: mostly synthetic single-persona corpora, small per-condition samples, semantic (not verbatim) agreement, and a self-training recipe that is not itself novel.

1. Introduction

Most systems that give a language model long-term memory treat memory as a tool the model uses: the model is the seat of behavior, and a retrieval store, scratchpad, or vector index feeds it context (RAG [1]; MemGPT/Letta [2]; Mem0 [3]; the memory stream of generative agents [4]). We invert the relationship. The model is an interchangeable process that reads from and writes to a memory substrate that is itself the durable object. Under this inversion the memory is not a cache of past prompts but a structured, deterministic account of a user that persists while the model reading and writing it is swapped out.

The inversion is only worth stating if it can fail a test, and there is a reasonable prior under which it should. Language models are not interchangeable extractors: a 1B and a 4B model, or a Gemma and a Qwen, differ in what they treat as salient, in their extraction and paraphrase biases, and in the world knowledge that bleeds into what they write down. If a system's account of a user were the sum of those model-specific choices — if it lived in the weights — then swapping the model across a 4× scale gap and a family change should visibly shift the account. That it does not is the finding; the apparatus below is built to make the non-shift measurable rather than assumed. If instead the account lives in the substrate, it should survive the swap. Most of this paper measures that from several directions — swapping the reader, swapping the writer, swapping the writer mid-stream, comparing the shape of the resulting self, and lesioning the substrate to check that agreement is not merely an artifact of shared input — under one methodological commitment: every claim is measured against a floor, a ceiling, or a null control, so a reported number sits between two things a reader can check.

We keep the systems claim (a model-independent account) firewalled from the interpretation (that this constitutes identity). The systems claim is what the numbers support and is what we ask reviewers to weigh; the identity reading is offered as a bracketed interpretation a reader can reject without losing the systems contribution. Neither the memory-augmented LLM nor the self-improvement loop of §9 is novel — the loop belongs to the STaR / Self-Instruct / ReST family [5–7], whose failure mode (model collapse [8]) and mitigation (filtering) are documented. Our contribution is the substrate framing and the apparatus that localizes where continuity lives.

Contributions, in order of how load-bearing they are:

1. A falsifiable apparatus that localizes continuity — write-side and read-side interchangeability with explicit floors, a lesion that isolates the substrate's contribution, and a perturbation battery with null controls (§4, §5, §6). This is the core.
2. A deterministic, model-independent memory substrate and the "model proposes / code decides / grounding non-negotiable" discipline that makes it auditable and reproducible (§2, §3).
3. Supporting properties: a field-like self realized as substrate structure (§7); a self-improvement loop whose gains live in the fabric, not the weights (§9); and structural abstention demonstrated at scale — zero confident false assertions over 1,200 adversarial items spanning 722 third-party personas, versus 154 for a bare RAG baseline with the same two readers, the weakest and strongest of the pilot both falling to zero (§8).
4. A bracketed reading of the above as substrate-realized identity, separable from the systems result.

This is a systems and agent-memory paper, and we write for that reader: the evaluation emphasis is on controls and reproducibility, and the design payoff — memory-centered agents whose continuity and trustworthiness are properties of the store rather than the model — is stated in §11.

2. The Substrate

The substrate is a graph. Each node is a persisted, resolved statement about the user, carrying its canonical claim text, the entities it mentions, the life-domains it touches, a reinforcement value, a temporal stamp, and a status. Entities are first-class and canonicalized (so "the 38", "the bus", and "the number 38 bus" resolve to one referent), and typed relations connect nodes (causes, motivates, before, contradicts). The distinction that makes the substrate an object rather than a transcript is between the Claim and the Node: a Claim is the transient thing a model proposes from raw experience and is never stored; a Node is what a deterministic resolver decides to commit. The model's output is a proposal, not a write.

The write path is deterministic and grounded. A proposed claim passes a firewall that classifies it (accept, rescue, reject, quarantine) against grounding rules; an entity linker canonicalizes its referents; then a pure resolver — no model in the loop — decides the mutation: create, dedup, reinforce, modify on an explicit change signal, mark a dispute when a contradiction should be held rather than resolved, or noop. Because the resolver is a pure function of (incoming claim, current graph), the same claim applied to the same graph produces a byte-identical mutation. That is what lets us hand one fixed snapshot to different models and compare what they do, and what makes every perturbation in §6 a measurement against an exact baseline.

The read path cannot fabricate into the store. Retrieval is deterministic and read-only; a single bounded reader call then either renders grounded prose or structurally abstains when the retrieved evidence is insufficient. The reader post-hoc-validates against the retrieved nodes rather than being trusted to be faithful, and nothing it emits is written back. Reading and writing are both gated by code, and the model, on either side, is a swappable renderer or proposer.

One rule governs all of the above and all of §7: the model proposes, deterministic code decides, and nothing ungrounded is committed. A set of load-bearing invariants, recorded as an explicit non-goals contract and enforced by the test suite, forbids the shortcuts that would collapse the inversion — no model-authored writes, no ungrounded significance, no silent deletion of sources. The point is not that the models are weak; it is that their outputs are treated as evidence to be adjudicated, which is what makes the resulting memory model-independent.

3. Method & Determinism

Perception and consolidation run on separate clocks. A fast perception loop turns incoming raw turns into candidate percepts with a small, cheap model; a buffer holds them; a slower consolidation pass runs the deterministic write path plus a decay step. Decoupling the two means responsiveness is independent of consolidation, and it localizes the only place a model touches state to a proposal the write path then adjudicates.

Determinism is what makes the substrate an object one can do science on. A fixed snapshot can be handed to different models and their behavior compared directly, since any difference is attributable to the model. Perturbations are exact: "delete this entity", "apply these sessions in a different order", "fork and rejoin" all produce a well-defined new state, so the delta from baseline is a real measurement rather than sampling noise. And the invariants are executable — a regression suite (600+ tests) pins the write path, the resolver's decision table, decay, and the field-like-self guards, so the determinism the argument relies on is checked continuously rather than asserted.

We measure interchangeability at three levels, and — this is the methodological spine of the paper — each number is placed between a floor and, where available, a ceiling, so it is interpretable. Read-side (§4.2) compares different reader models over one snapshot, with a shuffled-pair floor that isolates topical similarity from real agreement. Write-side (§4.1) compares the substrates different perceivers build, with a cross-persona floor (what two accounts of different people score) and a same-family temperature-0.7 ceiling

(what staying within one model family costs under realistic sampling, 0.92). Field-level (§4.3) compares the significance-weighted shape of the self. Throughout, the substrate's own determinism floor is 1.0 (byte-identical), and the null-fork / null-merge controls of §6 sit at their trivial values, so a positive result is legible against something a skeptic can check.

4. Localizing Identity: Interchangeability

If continuity lived in the model's weights, swapping the model would change the system's account of the user. It does not — on either boundary, and at the level of the self's overall shape. We report the write side first because it is the stronger and more discrete result, then the read side, then the field.

4.1 Write side

We feed the same raw experience through different perceiver models and compare the substrates that form. "Competent" is a criterion fixed in advance and independent of any agreement outcome: a perceiver is competent if it emits valid, structured, source-grounded claims at all. The base-0.5B model, for instance, produces zero usable structured claims over the corpus (§9) — a format-validity failure — and is excluded on that basis, not because it disagrees. Among three competent perceivers spanning a 4× parameter-scale gap and two families (gemma-3-1b, gemma-3-4b, qwen-4b), semantic claim agreement is 0.77–0.87 (Table 1; every claim similarity in this paper uses one sentence encoder [12], so read- and write-side numbers are on one scale).

Read that against its anchors, of which there are now three. Two accounts of *different* people score 0.29 (a cross-persona floor over four personas). Re-running a *single* perceiver over the same corpus at temperature 0 reproduces its claims exactly (1.00) — a determinism check, not an upper bound. The load-bearing ceiling is the third: a *same-family* pair (gemma-1b vs gemma-4b) run under realistic sampling (temperature 0.7, on a second, MSC-derived corpus) agrees at claim-semantic 0.92 — the non-hollow upper reference for what extraction agreement looks like when the model family is held fixed. (Being measured on a different corpus, it is a reference point rather than a strict same-corpus bound; we flag that rather than hide it.) The cross-family band (0.77–0.87) sits close beneath it: crossing a family boundary costs only slightly more agreement than staying within the family. Cross-model 0.8 against a 0.29 different-person floor and a 0.92 same-family ceiling is the write-side result in one line — two different models write substantially, though not identically, the same claim content about the same person.

Table 1. Write-side interchangeability (same encoder throughout; anchors: cross-persona floor 0.29, same-family temperature-0.7 true ceiling 0.92, determinism check 1.00).

pair	axis	entity-J	domain-J	claim-semantic
gemma-1b vs gemma-4b	scale	1.00	0.85	0.77
gemma-1b vs qwen-4b	scale + family	1.00	1.00	0.87
gemma-4b vs qwen-4b	family	1.00	0.85	0.83

The claim-semantic figure is the robust signal, and we lead with it deliberately. The discrete entity-agreement figure is directionally strong but fragile, and we flag its fragility ourselves. On the Maya corpus all three perceivers converge on the identical named-entity set {Van Ness, Civic Center, Pekoe}, an entity-Jaccard of 1.00 — but that is a three-element set on one persona, where Jaccard is high-variance and 1.00 is easy to reach. The very next section (§5) shows the same measurement dropping to 0.78 when one small perceiver misses one entity on a different run. We therefore treat entity-J as supporting evidence, not a headline, and report it with its across-persona variability: a four-persona replication holds claim-

semantic near 0.8 throughout (Maya and Priya strong, Theo and Marcus supported), while entity-J degrades gracefully as the one-off long tail of entities grows.

Replication on a third-party corpus. The personas above are ours; the same measurement on a corpus we did not author is the sharper test. We ran the *unmodified* harness on LoCoMo [15], a crowdsourced multi-session dialogue benchmark: we flattened one speaker's turns into a first-person account (Caroline, 30 turns) and ran the same three perceivers, with a second speaker (Jon) supplying a cross-*person* floor. The headline replicates: cross-perceiver claim-semantic is 0.79 (Table 2, same encoder), inside the synthetic 0.77–0.87 band and +0.43 above its cross-person floor of 0.36 — two different models write substantially the same account of a real person, and markedly less so about a different one. Jon's turns reproduce the pattern independently (claim-semantic 0.86–0.91). The fragile axis behaves as flagged: entity agreement degrades, because casual dialogue leads the entity linker to surface sentence-initial capitalized common words ("Gonna", "Wow", "It'll") that pass the verbatim source-grounding gate but vary across perceivers, whereas the genuine named entities ("LGBTQ", "Melanie") are shared. We diagnosed and partially closed this gap deterministically: a junk filter over the linker — a pinned word-frequency lexicon (excluding tokens at Zipf ≥ 5.0) plus contraction/pronoun exclusion, no model in the loop — raises mean entity-J from 0.43 to 0.52 (per-pair: 0.31⟶0.40, 0.57⟶0.67, 0.42⟶0.50) while leaving claim-semantic and domain agreement unchanged, confirming the failure was extraction noise, not a divergence of the account. The honest residual: moderate-frequency common words below the Zipf cutoff still leak, and cross-perceiver nickname canonicalization ("Mel" vs "Melanie") remains open — a semantic entity-name merge is the identified next step. Domain-J holds at 0.81. The reading is clean: on non-synthetic text the *semantic* account survives model swaps, while *discrete entity canonicalization* is corpus-quality-dependent, degrades on open-domain chat, and is partially recoverable with deterministic filtering — which localizes exactly which part of the substrate is robust and which is input-sensitive.

Table 2. Third-party replication (LoCoMo, Caroline; same encoder; cross-person floor 0.36; entity-J after the deterministic junk filter, pre-filter values in parentheses).

pair	axis	entity-J	domain-J	claim-semantic
gemma-1b vs gemma-4b	scale	0.40 (0.31)	0.71	0.73
gemma-1b vs qwen-4b	scale + family	0.67 (0.57)	0.86	0.83
gemma-4b vs qwen-4b	family	0.50 (0.42)	0.86	0.81

4.2 Read side

Retrieval is deterministic, so the read side asks whether different reader models, given the same fixed snapshot, give the same answer. At the surface they do not: token overlap between two readers' answers is low (0.18–0.20 Jaccard), and a small reader can be barely self-consistent (gemma-1b agrees with itself only 0.28 of the time). But surface wording is the wrong metric, and we show so with a control rather than an assertion. Measured semantically on the same sentence encoder as the write side, matched-pair answer similarity is 0.69–0.77, against a shuffled-pair topical floor of 0.33 — the same answers re-paired to the wrong queries, which isolates mere shared vocabulary. A permutation test (2000 shuffles) places the matched mean far outside the shuffled distribution: $p < 0.001$ for both pairs. Two different readers give the same answer because it answers the asked question, not because they share the corpus's vocabulary. This replicates across two model pairs, including a 4× scale and cross-family pair, and the readers agree on when to abstain (0.78–0.94).

Table 3. Read-side interchangeability (sbert, same encoder as Table 1). Surface token overlap is low; matched semantic similarity sits far above the shuffled-pair floor.

pair	axis	surface token-J	matched	shuffled floor	perm p	abstain-agree
qwen-4b vs gemma-4b	family	0.20	0.77	0.34	<0.001	0.94
gemma-1b vs qwen-4b	scale + family	0.18	0.69	0.33	<0.001	0.78

4.3 Field level

Beyond individual claims, we measure the shape of the self as a significance-weighted field (mass, centroid, dispersion, concentration). Across gemma-1b versus qwen-4b, and across a mid-stream model transplant (the perceiver swapped halfway through the stream), the field keeps the same centroid — the entity the self orbits: the cross-model deltas are small (dispersion 0.11, concentration 0.15) and the transplant deltas smaller (0.04, 0.09); only the mass differs, as a more verbose perceiver grows a larger self (mass ratio 0.53–0.68). The centroid is the discriminating quantity, and it does not saturate: it is a named entity, and different personas have disjoint entity sets (§5: Maya's {Van Ness, Civic Center, Pekoe} versus Theo's {Marrow, Hector, Lin}), so two accounts of different people share a centroid essentially never — "same centroid" across a model swap is therefore a real coincidence to explain, not a foregone one. The self orbits the same point regardless of which model wrote it.

Table 4. Field-level self-shape across a model swap and a mid-stream transplant.

comparison	same centroid	mass ratio	dispersion Δ	concentration Δ
gemma-1b vs qwen-4b (cross-model)	yes	0.53	0.11	0.15
single vs mid-stream transplant	yes	0.68	0.04	0.09

The simplest account of §4 is that the user's identity, as the system represents it, is carried by the substrate and is invariant to the model. A skeptic may prefer "different competent models extract overlapping facts." The next section is built to separate those two readings.

5. A Non-Substrate Control (The Lesion)

The interchangeability results invite one sharp objection: if two perceivers agree, is that the substrate, or would any store agree because both models read the same text? We answer it by ablation. We hold perception identical and remove only the substrate's deterministic consolidation — entity canonicalization/merge, claim resolution/dedup, and the grounding firewall. The substrate store is the full pipeline; the naive-append store is the same perceiver output kept raw (every claim, every surface entity string with exact-match dedup only, no canonical merge). We then compare the stores that three perceivers (gemma-1b, gemma-4b, qwen-4b) build, with the same metrics.

Table 5. Lesion — cross-perceiver agreement, substrate vs naive-append (identical perception).

cross-perceiver agreement	substrate	naive-append	Δ
entity-Jaccard	0.78	0.10	+0.68
domain-Jaccard	0.90	0.30	+0.60
claim-semantic	0.88	0.89	−0.01

The result is precise, and more informative than a uniform win. The structured account collapses without the substrate: the three perceivers' naive entity sets agree at 0.10 — the raw store retains divergent, uncanonicalized surface tokens ("i", "it", "38 bus" in one model, "van ness stop" in another, "cat" in a third) — where the substrate canonicalizes all three to the same set (0.78 here; 1.00 in the dedicated §4.1 run, the difference being one small perceiver missing one entity on this pass). Domain structure behaves the same way (0.90 versus 0.30). But raw claim similarity is unchanged (0.88 versus 0.89), because the claims are the same perceiver sentences either way. So the lesion localizes the substrate's contribution exactly: it is the canonical who-and-what the account is about — the entity and domain structure — that a non-substrate memory fails to preserve across a model swap, not the surface text (which any store keeps). This is the externally-legible anchor for §4: the part of the write-side result that carries the account is a substrate property, and a store lesioned of that machinery does not reproduce it.

6. Identity Under Perturbation

Personal-identity philosophy supplies the hard cases — fission, fusion, amnesia, gradual change, transplant [11]. Because the substrate is deterministic, we turn each into a measurement, and pair each with a null control that discriminates. We ran the full battery with the field-like self of §7 active, to check that a richer self is not a more fragile one. Perceiver gemma-3-1b; cross-family transplant qwen-4b.

Table 6. Perturbation battery, each paired with a discriminating null control.

Perturbation	Thought experiment	Result	Null control
Swap	transplant (read & write)	write-side claim-sem ~0.8, entity-J 1.0 (§4); read-side matched 0.7–0.8, p<0.001	shuffled-pair floor 0.33; cross-persona floor 0.29
Age (14 sessions)	gradual change	determinism 1.0, order-invariance 1.0; cross-model claim-sem 0.82, transplant 0.91	determinism floor 1.0
Fork	fission	continuity 0.67; experience-driven divergence (A-vs-B 0.82)	null-fork 1.0
Merge	fusion	same-person re-integration, dedup 0.64	different-person contamination 0.0, 0 chimera nodes
Corrupt	amnesia	graceful to 50% deletion; heals by re-exposure (recovery 1.0, 0 id-collisions)	targeted > random sensitivity
Hot-swap	transplant, live	gemma-1b ⟶ qwen-4b mid-session: 18/18 pre-swap nodes carried	—

The null controls are the anchor: a real fork diverges (0.82) where a null fork does not (1.0); a different-person merge contaminates at 0.0. Identity has load-bearing structure — deleting an identity-central entity damages the account more than deleting a random one. Running the full field-like self did not make the self more fragile: every perturbation strong before is strong now, and the write path is byte-identical (determinism and order-invariance 1.0). These are the classical thought experiments made empirical; the substrate's account of a person survives fission, fusion, amnesia, gradual change, and transplant.

7. The Field-Like Self (Capability Description)

This section describes what the substrate carries beyond a fact list. Unlike §4–§6 it is a capability description, not a controlled result, and we do not offer it as evidence for the thesis; we include it because the perturbation battery (§6) was re-run with all of it active, so a reader needs to know what "active" means. Every property is a side-table or typed edge over the existing graph, added under the same rule — the model proposes, code decides, nothing ungrounded is committed — so a richer self is not a more credulous one.

A perceiver may propose, for a claim, a rationale ("why this matters") and a cause (the source clause it is caused-by); a grounding guard admits a rationale only if its content traces to the source and adds content beyond the claim, and a cause only on a real causal connective plus span containment. Run on a tiny perceiver over a 26-turn corpus, the guard kept 8 of 47 proposed rationales and 3 of 33 proposed causes, rejecting every fabrication and every reworded restatement. Grounded causes and rationales are promoted into typed causes/motivates edges (with deterministic before edges from event times), raising relational density from 0.10 to 0.31 edges per node, so an opt-in multi-hop retrieval can answer "why" questions that plain retrieval cannot. A contradiction is not always an update: with an explicit change signal ("now",

"no longer") a claim supersedes, but a bare opposition is held — both claims stay live, linked by a contradicts edge, and both surface at read time. A background scheduler runs opportunity-scored consolidation (freeze a meaningful memory against decay; propose an evidence-to-pattern reflection past overreach and grounding guards; compress cold clutter under anti-destruction guards that never fold a meaningful, frozen, or sole-domain node). Forgetting is three-axis: core and frozen nodes are immutable, reinforcement decays, and grounded significance damps the decay rate. Finally, a constitution of the system's values is seeded as immutable core nodes, so the values, like the rest of the account, are carried by the substrate and survive a model swap. The precision of these guards is reported as counts, not as a benchmark; a held-out evaluation of the significance guard against human labels is future work.

8. Structural Abstention (Supporting Property)

A separate and smaller question: is the account trustworthy, and is that trust a property of the substrate or of the model? On a 40-question benchmark over real development history, the read path produced zero confident false assertions across four runs (160 Q-A pairs) and abstained correctly on 15 of 15 out-of-scope questions. But a modern instruct model abstains fairly well on its own when context obviously lacks the answer, so that alone does not isolate the substrate's contribution.

We therefore tested under adversarial retrieval load, in two stages: a small five-reader pilot, whose statistical limits we report in full, and then a scaled run designed to remove exactly those limits.

Pilot. 12 questions whose answer is absent but for which a plausible distractor is retrievable (over-generalization, attribute/temporal swaps, false premises), asked of a bare RAG (sbert retrieval plus a neutral prompt) and of the same model inside Fireweed's structural pipeline, swept across five readers (three families, 1B–4B). Answers were classified by an LLM judge, whose reliability we quantify rather than assert: on a 24-item stratified subset labeled by the author with a rule fixed in advance, judge–human agreement was 92% and Cohen's $\kappa = 0.75$ (substantial; single annotator, a limitation).

Table 7. Pilot — confident false assertions under adversarial load, per reader (raw counts; per-cell n=12 is too small for per-reader inference — see the pooled analysis below).

reader	bare RAG	inside Fireweed
gemma-3-1b	6/12	4/12
gemma-3-4b	4/12	3/12
qwen3-1.7b	4/12	1/12
qwen3-4b	1/12	0/12
liquid-1.2b	3/12	1/12
pooled	18/57	9/57

The pilot's quantitative picture is directional but not significant, and we are careful not to overstate it. Fabrication fell for all five readers tested (6⟶4, 4⟶3, 4⟶1, 1⟶0, 3⟶1), a pooled reduction of about 0.16 (halving the count, 18 to 9). But the items are not independent — the same twelve questions recur across all five readers — so a pooled test that assumes independent pairs (McNemar) overstates the evidence. Once we resample at the cluster level, bootstrapping over questions and readers, the 95% CI on the reduction crosses zero ($[-0.05, 0.38]$), and a reader-level sign test is two-sided $p = 0.06$. On the pilot alone we would not claim a significant rate reduction. The pilot's three weaknesses are explicit: one author-written persona, twelve clustered questions, one human annotator.

At scale. We then removed each weakness by construction. The scaled benchmark generates 1,200 adversarial items over 722 distinct third-party personas (drawn from the MSC corpus [16], not authored by us), classified with a three-way rubric fixed in advance (ASSERT / STRUCTURAL-ABSTENTION / HEDGE) by an ensemble judge of three local models rather than one, whose agreement against the human-labeled slice is $\kappa = 0.88$ (raw agreement 96%). Two readers spanning the pilot's capability range — its weakest (gemma-3-1b) and strongest (qwen3-4b) — answered every item in both configurations: 2,400 answer opportunities per configuration. Inside Fireweed's structural pipeline they produced **zero** confident false assertions; the bare-RAG configuration produced **154** — 117 from the weak reader, 37 from the capable one, and both fall to zero inside the substrate. With 722 persona clusters, the clustering concern that undercut the pilot no longer applies — there is no resampling of these data under which zero is compatible with 154. We state the claim exactly as wide as the design licenses: under adversarial memory load — questions whose answers are absent but for which plausible distractors are retrievable — the substrate structurally refuses to fabricate, independent of reader capability. It is *not* a claim that a reader can never misstate a fact the graph does contain; that residual risk is the reader's, bounded by post-hoc validation against retrieved nodes, and is not measured here.

The scaled result and the provenance guarantee together carry the section. Every substrate answer traces deterministically to specific grounded nodes, and abstention is a structural consequence of insufficient evidence rather than a behavior we hope the reader exhibits; a RAG answer is reader-synthesized and unverifiable. What remains concentrated in the authors is the *generator* and the *rubric*: the adversarial item templates and the judge rule are author-written, and the human validation slice is one annotator's (the ensemble $\kappa = 0.88$ is against those labels). The mitigation is the released bundle — the judged prose and labels ship with the paper, so an independent party can re-label, recompute κ, and re-run the statistics. With the scaled run, §8 graduates from a directional pilot to a substantive property, though we still present it as supporting the thesis (trustworthiness as a substrate property) rather than as the thesis itself.

9. Closing the Loop (Supporting Property)

The substrate's deterministic mutations are also a clean training signal, which lets us close a self-improvement loop — the perceiver harvests its own well-formed, grounded claims and fine-tunes the operator — and ask where the improvement accumulates. The loop mechanism is prior art [5–8]; the substrate framing is what lets us localize the answer, via two controls detailed in Appendix B. Naive iteration does not beat training once on the union of the same data, so nothing compounds in the weights; but the fine-tune does confer a capability the base lacks — structured, grounded perception (the base emits 0/12 usable claims, the trained model 12/12, which is exactly the format-validity gate §4.1 relies on). The fine-tune is a removable adapter over a frozen base [9,10], so forgetting across iterations is impossible by construction, and a dual-eval harness confirms no catastrophic forgetting at the canonical 3B scale (Appendix B). We keep this brief and clearly supporting: the runs are unseeded and the samples small, and the argument of §4–§6 does not rest on it. The one-line synthesis is that any compounding lives in the fabric, invariant to the model, not in the weights.

10. Related Work

Memory-augmented LLMs. Retrieval-augmented generation [1], MemGPT/Letta [2], Mem0 [3], and the memory stream of generative agents [4] give a model durable context but leave the model as the seat of behavior. Our inversion makes the store the durable object and the model the tenant, and asks the interchangeability question those systems do not pose.

Knowledge-graph construction and entity resolution. The substrate is, mechanically, deterministic knowledge-graph construction from text with first-class entities and typed relations, and its canonicalization

step is entity linking and coreference resolution. We inherit that framing and the evaluation habits (entity/relation agreement) of those fields; what is new here is treating the *stability of the constructed graph across the constructing model* as the object of study.

Belief revision and truth maintenance. Fireweed's held-contradiction behavior — keep both sides of an opposition live and marked, rather than collapse them — is a lightweight, grounded analogue of truth-maintenance and belief-revision systems (e.g. de Kleer's assumption-based TMS [14]) and the AGM belief-revision tradition. We do not implement a full logic; we borrow the stance that contradiction is information to be represented, not an error to be erased.

Self-improvement. STaR [5], Self-Instruct [6], ReST [7], and rejection-sampling fine-tuning, with model collapse [8] as the known failure and filtering as the known fix, are the family our §9 loop belongs to; our addition is the substrate framing and the weight-versus-fabric controls, over a low-rank adapter [9,10].

Evaluation. Hallucination and abstention in RAG is the backdrop for §8; LLM-as-judge evaluation is the method we use there, and we follow its emerging norm of reporting judge–human agreement. Semantic agreement throughout uses sentence-embedding similarity [12] and calibration the Brier score [13]; the personal-identity thought experiments are Parfit's [11].

11. Discussion & Limitations

What the results support is a systems claim: a deterministic memory substrate preserves a system's account of a user — the canonical entities and domains, the self's shape, and its behavior under perturbation — when the reading and writing models are swapped across a 4× scale gap, two families, time, and the classical perturbations, and a store lesioned of that machinery does not. For the memory-centered agent designer, the payoff is concrete: continuity and trustworthiness become properties you get from the store, so you can swap or upgrade the model without migrating the user's identity; you can audit what the agent believes because every belief is a grounded node; and under adversarial memory load the system refuses to fabricate as a structural matter (§8), not as a behavior of whichever model is mounted. The stronger reading — that this is identity, substrate-realized rather than weight-bound — is an interpretation we find natural but keep bracketed; a reader can reject it and lose nothing of the systems contribution.

A systems note on cost: because retrieval is a deterministic query over the graph rather than long-context token dumping, the substrate is also *cheaper* to read. On a 50-episode LongMemEval [17] slice, answer accuracy was at **statistical parity** with a RAG baseline (Fireweed 0.42 vs 0.48, approximate contains-match) — the difference is well inside binomial noise at n = 50 (±~0.14) — while mean end-to-end latency was **~2.5× lower** (6.71 s vs 16.53 s) with **no pathological tail** (max 13.9 s vs 241 s), consistent with a 10-episode slice (0.40/0.40) showing no accuracy cliff across scale. Fifty episodes supports direction more than a definitive scale claim; a validated 500-episode run behind a perceiver-competence gate is the remaining external-validity step (below).

Limitations, stated plainly. (1) Corpora: most of the *interchangeability* evidence rests on a small number of synthetic personas, mitigated with a four-persona write-side replication, a 14-session longitudinal study, and a third-party replication on LoCoMo [15] (§4.1) — on which the *semantic* interchangeability headline holds (claim-sem 0.79, +0.43 over a cross-person floor) while *discrete entity canonicalization* degrades on open-domain chat (entity-J 0.52 after a deterministic junk filter, up from 0.43 before it) — an honest, diagnosed, partially-closed gap rather than a divergence of the account. Cross-perceiver nickname canonicalization ("Mel"/"Melanie") remains open. The interchangeability corpora are still not a large naturalistic *population* (two LoCoMo speakers, one focal); the §8 abstention result, by contrast, now spans 722 third-party personas. (2) The pilot abstention effect (Table 7) is, on its own, not significant under cluster-robust resampling; the scaled run (1,200 items, 722 persona clusters, two readers, 0 vs 154) is what carries §8, and its

scope is exactly structural non-fabrication on unanswerable items — not reader infallibility on answerable ones. (3) Author-constructed evaluation: the §8 adversarial *generator* and judge rubric are author-written and the human validation slice is one annotator's (ensemble judge $\kappa = 0.88$ against it); the released judged prose and labels let an independent party re-label and recompute. (4) Competent-perceiver scope: interchangeability holds among models that pass a pre-fixed format-validity gate (§4.1), not "any model." (5) Semantic, not identity: cross-model claim agreement is ~0.8 by embedding similarity, not verbatim; what carries continuity is the structure (entities and domains are the robust, discrete signals), not the exact wording. (6) Single training runs: the loop results (§9, Appendix B) are unseeded. (7) Scale: end-to-end evidence on long-horizon third-party benchmarks is at the 50-episode scale (statistical parity on accuracy within binomial noise, ~2.5× lower latency with no tail); a 500-episode run remains the open gate. An earlier 500-episode attempt was invalidated — an unvalidated perceiver produced near-empty substrates on 79% of episodes — and is disclosed here as a non-result, folded into no claim; a perceiver-competence sanity gate now precedes any full run. An early MemoryAgentBench conflict-split probe was inconclusive on accuracy under a capped multi-hop budget (while still showing fewer ensemble-judged fabrications, 12 vs 17). (8) Prior art: the self-training recipe is not novel. (9) Reproducibility: our reproducibility claim is only meaningful if others can exercise it; see Availability.

12. Conclusion

Identity, in this system, is carried by a deterministic memory fabric and is largely invariant to which model reads and writes it — across scale, family, time, and the classical perturbations — and a store without the fabric's machinery does not preserve it. Trustworthiness follows the same pattern: under adversarial memory load the fabric refuses to fabricate structurally, at scale, whichever reader is mounted. The fabric is the product; the model is the tenant.

Availability & Reproducibility

The Fireweed implementation is not open-sourced (the deterministic resolver is the contribution), but a public **verification bundle** that does not depend on it is released at `https://github.com/Starksood/Fireweed_Fabric`. It contains: (a) the committed benchmark result JSONs, including the raw per-query model answers and the judged prose, so every claim rests on inspectable outputs rather than prose; (b) the scripts that recompute the statistics from those raw outputs — the shuffled-pair read-side floor and permutation test, the write-side cross-persona floor and temperature-0.7 true ceiling, the cluster-bootstrap and reader sign test for the §8 pilot, the scaled-run counts and ensemble-judge agreement, the judge–human agreement, and the lesion deltas; and (c) `make_paper_figures.py`, which regenerates Tables 1–7 from the JSONs. A reviewer cannot re-run perception (that needs the system), but can independently verify every floor, ceiling, null control, statistic, and the lesion — the claims that carry the argument. A manifest lists the exact files with SHA-256 digests.

References

[1] Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." arXiv:2005.11401; NeurIPS 2020. [2] Packer et al. "MemGPT: Towards LLMs as Operating Systems." arXiv:2310.08560, 2023. (Letta.) [3] Chhikara, Khant, Aryan, Singh, Yadav. "Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory." arXiv:2504.19413; ECAI 2025. [4] Park et al. "Generative Agents: Interactive Simulacra of Human Behavior." arXiv:2304.03442; UIST 2023. [5] Zelikman et al. "STaR: Bootstrapping Reasoning with Reasoning." arXiv:2203.14465; NeurIPS 2022. [6] Wang et al. "Self-Instruct: Aligning Language Models with Self-Generated Instructions." arXiv:2212.10560; ACL 2023. [7] Gulcehre et al. "Reinforced Self-Training (ReST) for Language Modeling." arXiv:2308.08998, 2023. [8] Shumailov et al. "The Curse of Recursion: Training on Generated Data Makes Models Forget." arXiv:2305.17493, 2023; published as "AI

models collapse when trained on recursively generated data," Nature 631, 2024. [9] Hu et al. "LoRA: Low-Rank Adaptation of Large Language Models." arXiv:2106.09685; ICLR 2022. [10] Dettmers et al. "QLoRA: Efficient Finetuning of Quantized LLMs." arXiv:2305.14314; NeurIPS 2023. [11] Parfit. *Reasons and Persons.* Oxford University Press, 1984. [12] Reimers & Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." arXiv:1908.10084; EMNLP 2019. [13] Brier. "Verification of Forecasts Expressed in Terms of Probability." Monthly Weather Review, 78, 1950. [14] de Kleer. "An Assumption-Based TMS." Artificial Intelligence 28(2), 1986. [15] Maharana et al. "Evaluating Very Long-Term Conversational Memory of LLM Agents." arXiv:2402.17753; ACL 2024. (LoCoMo dataset.) [16] Xu, Szlam, Weston. "Beyond Goldfish Memory: Long-Term Open-Domain Conversation." arXiv:2107.07567; ACL 2022. (Multi-Session Chat corpus; source of the 722 §8 personas.) [17] Wu et al. "LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory." arXiv:2410.10813; ICLR 2025.

Appendix A: Superseded Metrics & Provenance

Kept for scientific hygiene; not part of the argument.

- **Read-side "90.73%."** An earlier internal figure for read-side interchangeability was a within-model self-consistency number (one reader agreeing with itself across replicates), not a cross-model one. It is superseded by the floor-controlled cross-model result in §4.2 (matched 0.69–0.77 vs 0.33 shuffled floor, permutation $p < 0.001$).
- **Stage-3 "92.3%."** Likewise a within-model number; not used in this paper.
- **σ reporting.** An earlier draft described the read-side lift as "≈15–17σ above the floor." That described the matched mean's distance from the mean of the shuffled-pairing distribution and implied a test it did not run; §4.2 reports a permutation p-value instead.
- **Abstention "McNemar p = 0.049."** An earlier draft reported a pooled McNemar test as significant. That test assumes independent pairs, which is violated: the 57 items are 5 readers × 12 shared questions. Under cluster-robust resampling the effect is not significant (§8); the McNemar figure is retained only as the (invalid) pooled number and should not be cited.
- **Two embedders.** An earlier draft measured read-side semantics with a different embedder than the write side. All claim/answer similarities are now on one sentence encoder (sbert, all-MiniLM-L6-v2); the earlier read-side numbers (matched 0.80–0.85, floor 0.61) were on a higher-baseline embedder.
- **Hollow ceiling (V4).** Draft V4 anchored the write side with a temperature-0 same-perceiver rerun (1.00) and flagged it as hollow. Superseded in V5 by the same-family temperature-0.7 ceiling (claim-sem 0.92: gemma-1b vs gemma-4b under sampling, on an MSC-derived corpus); the 1.00 figure is retained only as a determinism check.
- **LoCoMo entity-J 0.43 (V4).** Pre-junk-filter value. Superseded by 0.52 (per-pair 0.40/0.67/0.50) after the deterministic Zipf-frequency + contraction/pronoun filter (§4.1); claim-semantic and domain-J unchanged by the filter. Both values remain in Table 2 for provenance.
- **Abstention pilot as sole §8 evidence (V4).** V4′s §8 rested on the 5-reader × 12-question pilot (directional, cluster-CI crossing zero). Superseded as the section's primary evidence by the scaled run (1,200 items / 722 MSC personas / ensemble judge $\kappa = 0.88$: 0 vs 154); the pilot is retained as Table 7 with its original honest statistics.

Appendix B: Closed-Loop Detail (§9)

Single unseeded runs; small samples. Reported as supporting evidence only.

- *Setup.* Removable LoRA adapter over a frozen base [9,10]; dual-eval on three suites — in-domain held-out (n=4), out-of-distribution perception (n=12), general-knowledge canary with a JSON-leak tripwire (n=12).

- *0.5B pilot.* Held-out loss 3.24 ⟶ 0.92; claim-JSON validity 0/3 ⟶ 3/3.
- *Canonical 3B* (Qwen2.5-3B, 4-bit QLoRA). Base ⟶ adapter: OOD valid-JSON 1/12 ⟶ 12/12, claim-F1 0.90 on unseen personas, general hit-rate 12/12 unchanged, JSON-leak 0 — no catastrophic forgetting.
- *Sequencing control (0.5B, matched gradient steps).* Training once on the union met or beat iterative self-training on every task metric (union 12/12 valid-JSON, claim-F1 1.0; iterative 10/12, 0.83): naive iteration does not compound in the weights. At 3B the comparison is confounded (iterative ran ~2× the optimizer passes) and inconclusive.
- *Bootstrapping control.* On identical raw text the base produced 0/12 usable self-generated claims and the iterated model 12/12 — a format-validity difference (the base fails to emit the structured schema), which is the gate §4.1 relies on, not a reasoning gain.