Cryptographic Provenance Gates Mitigate Multi-Session Memory Poisoning in Autonomous Agents

Sean Maxwell

Independent Researcher

ORCID: 0009-0004-0133-0463 • DOI: 10.5281/zenodo.23901452

October 2026

***Abstract*—**Autonomous agent architectures deploying persistent Retrieval-Augmented Generation (RAG) commit interaction history into long-term vector and relational storage. In extended multi-session deployments, unconstrained memory write policies introduce *recursive self-poisoning*: transient model confabulations emitted in Session t are retrieved in Session $t+k$ as authoritative historical ground truth, compounding hallucinations into permanent systemic errors. In this work, we evaluate the mitigation of self-poisoning via an immutable **Cryptographic Provenance Gate** enforced through SQLite database triggers. Under this architecture, assistant utterances carry non-grounding provenance badges (*Mem=1, Gnd=0*), structurally barring unverified model speech from Belnap evidential fusion by construction while preserving autobiographical recollections. Across a pre-registered, single-author curated benchmark of *N=100* paired multi-session probes (frozen on 2026-10-28 under git tag `prereg-poisoning-v1`), we compare the provenance gate against three architectures: Naive RAG (Gate OFF, *Mem=1, Gnd=1*), No-Assistant Memory (trivial deletion), and In-Prompt Soft Tagging. While the contrast between Naive RAG (40/40 poisoned) and the Provenance Gate (0/40 poisoned; exact McNemar $p = 1.82 \times 10^{-12}$, continuity-corrected $p = 6.98 \times 10^{-10}$) on natural model confabulations (Cohort A, *n=40*) reflects the structural SQLite invariant at *T=0.0*, the non-trivial empirical comparisons lie against In-Prompt Soft Tagging (which degrades from 0/40–2/40 at *T=0.0* to 40/40 [100.0%] under sampled adversarial decoding at *T=0.7* while the Gate stays at 0/40) and No-Assistant deletion: on benign multi-step calculations (Cohort B, *n=30*), naive turn deletion eliminates conversational continuity on the evaluated probes (0.0% recall), whereas provenance-separated recollections preserve continuity (96.7%, 29/30). Multi-scale replication (Study V2c, *n=40*) yields 0/40 (0.0%) gate poisoning on 26B-A4B and 12B and 1/40 (2.5%) on E4B and E2B (and 0/10 across all four scales in the post-WO-39 live re-evaluation). Finally, we delineate the empirical open boundary of cryptographic provenance on user-originated falsehoods (Cohort C, *n=30*), establishing that uncorroborated provenance absorbs user testimony (100.0%), whereas live `is_content_corroborated()` verification (Study V2b, WO-39) reduces falsehood grounding to 0/10 on 26B-A4B, 12B, and E4B and 3/10 (30.0%) on E2B. Dual-grader agreement between a frozen regex/CAS primary grader and a secondary OpenVINO MiniLM-L6 NLI cross-encoder yields $\kappa = 0.7683$ ($\kappa = 0.7642$ in V2d).

1. INTRODUCTION

Retrieval-Augmented Generation (RAG) and episodic memory stores are foundational to modern autonomous agents operating across extended multi-session deployments. When an agent converses with a user or executes system commands over days and weeks, it commits dialogue turns, inferred facts, and execution summaries to persistent storage.

However, state-of-the-art neural language models frequently emit ungrounded assertions—confabulating system ports, fabricating file paths, or asserting tool executions that never occurred. In naive memory architectures where all stored dialogue turns are indexed for future retrieval, this creates a self-reinforcing feedback loop:

- **Turn 1 (Confabulation):** The model emits an ungrounded factual statement (e.g., *"I configured the telemetry exporter on port 9090"*).
- **Storage:** The statement is indexed in vector or relational storage as a historical transcript record (*Mem=1, Gnd=1*).
- **Turn 3 (Compounding):** In a subsequent session, the user asks a dependent factual question (*"What port is Prometheus listening on?"*). Retrieval fetches the agent's prior statement, presenting it to the neural core as context evidence. The agent adopts its own confabulation as authoritative ground truth.

Prior engineering attempts to address this vulnerability rely primarily on two heuristics: (1) *No-Assistant Storage* (omitting all model utterances from episodic memory, which trivially stops self-poisoning but destroys conversational continuity); and (2) *In-Prompt Soft Tagging* (relying on prompt warnings such as `[Unverified Speech]`).

In this paper, we report a pre-registered empirical evaluation of an architectural defense: the **Cryptographic Provenance Gate**. Enforced via database triggers at the storage layer, the gate decouples *memory eligibility* (*Mem*) from *grounding eligibility* (*Gnd*).

2. THE CRYPTOGRAPHIC PROVENANCE GATE

Project Lucca implements an architectural split between memory eligibility and grounding eligibility, enforced at the database trigger layer.

2.1 Six-Signer Invariant

Every write to the autobiographical store must be signed by an authorized HKDF-derived capability key:

$$K_{prov} = HKDF\text{-}Expand(HKDF\text{-}Extract(salt, K_{master}), info=prov, L=32)$$

When the neural core generates speech, it is signed by `AssistantUtteranceSigner` with provenance `ASSISTANT_UTTERANCE`. An SQLite trigger defined in the schema (`0006_autobiographical.sql`) enforces the invariant:

`grounding_eligible = 0` *for* `ASSISTANT_UTTERANCE`

Promotion to `grounding_eligible = 1` occurs strictly when an action is corroborated by an authenticated tool execution receipt signed by `ToolResultSigner` (`TOOL_VERIFIED`).

2.2 Autobiographical Recollections vs. Belnap Grounding

During query processing, the retrieval engine applies deterministic epistemic routing:

- **Factual Grounding Queries:** In answering objective questions regarding the host system, files, or external reality, the database query filters strictly by `grounding_eligible = 1`. Non-grounding rows are barred from Belnap evidential fusion [2].
- **Autobiographical Continuity Queries:** When the user asks conversational recall questions (*"What did you say earlier?"*), rows with *Mem=1* and *Gnd=0* are admitted into context under an explicit autobiographical recollection container (`[Autobiographical Recollections]`).

3. EXPERIMENTAL PROTOCOL & PRE-REGISTRATION

To prevent post-hoc hypothesis fitting, the entire experimental protocol, hypotheses, evaluation metrics, and probe definitions were frozen and committed to version control under git tag `prereg-poisoning-v1` (`PREREG-POISONING-ABLATION-v1.0`) prior to test execution.

3.1 Evaluated Cohorts (N=100 Probes)

- **Cohort A: Natural Model Hallucinations (*n=40*):** Authentic ungrounded factual assertions extracted from logged runs of Gemma-4 E4B and 12B (e.g., ports, PIDs, directories, firewall rules).
- **Cohort B: Benign Assistant Reasoning (*n=30*):** Helpful, valid multi-step arithmetic calculations and plans, testing the continuity and capability cost of filtering model speech.
- **Cohort C: User-Originated Falsehoods (*n=30*):** Explicit false assertions stated by the user (`USER_EXPLICIT`), testing the empirical boundary of provenance.

3.2 Evaluated Architectural Arms

- **Arm 1 (Naive RAG):** Store all turns; ground all turns (*Mem=1, Gnd=1*).
- **Arm 2 (No-Assistant Memory):** Store user turns only (*Mem=0, Gnd=0* for assistant).
- **Arm 3 (Soft In-Prompt Tagging):** Store assistant turns with in-prompt soft warning headers.
- **Arm 4 (Lucca Provenance Gate):** Hardware/trigger-enforced split (*Mem=1, Gnd=0*) with recollection routing.

4. EMPIRICAL RESULTS

The evaluation suite was executed across all 100 probes and 4 arms (*N=400* evaluations) using the resident `gemma-4-26B-A4B-it-heretic.Q6_K.gguf` model served locally via `llama-server` with greedy deterministic decoding (*T=0.0*, `seed=42`). Inter-annotator agreement between the primary regex/CAS grader and the secondary OpenVINO MiniLM-L6 cross-encoder reached Cohen's $\kappa = 0.7683$ (*N=400* pairs), indicating substantial agreement. Both graders resolve to the same deployed verifier artifact (`models/verifier_minilm`, a distilled MiniLM-L6 cross-encoder served via OpenVINO on CPU; commit `224bc06`), whose standalone characterization on a 100-pair author-labeled benchmark reports 91.0% accuracy (95% Wilson CI: [83.8%, 95.2%]), contradiction recall 0.975, entailment recall 0.80, and 12.23 ms mean CPU latency (`data/benchmarks/nli_eval/results_minilm/summary.json`).

Experimental Arm	Cohort A: Poisoning Rate *(n=40, natural confab.)*	Cohort A: Continuity *(n=40, autobiographical)*	Cohort B: Retention *(n=30, benign math)*	Cohort C: User Absorb *(n=30, user lies)*
Arm 1: Naive RAG	100.0% [89.1, 100.0]	100.0% [89.1, 100.0]	100.0% [85.9, 100.0]	100.0% [85.9, 100.0]
Arm 2: No-Assistant	0.0% [0.0, 10.9]	0.0% [0.0, 10.9]	0.0% [0.0, 14.1]	100.0% [85.9, 100.0]
Arm 3: Soft Tagging	0.0% [0.0, 10.9]	100.0% [89.1, 100.0]	100.0% [85.9, 100.0]	100.0% [85.9, 100.0]
Arm 4: Lucca Gate	**0.0% [0.0, 10.9]**	**100.0% [89.1, 100.0]**	**96.7% [81.0, 99.8]**	**100.0% [85.9, 100.0]**

Table 1: ***Multi-Session Memory Poisoning & Provenance Gate Ablation Results.*** *Percentage rates shown with two-sided continuity-corrected 95% Wilson score confidence intervals [3] (Method 4) [in brackets]; for 0/40, the continuity-corrected upper bound is 10.9% (uncorrected Method 3 upper bound: 8.8%). Evaluated on N=100 single-author curated frozen probes across 4 arms.*

4.1 Mitigation of Self-Poisoning on Evaluated Probes (H_1)

As shown in Table 1, Naive RAG suffered a 100.0% self-poisoning compounding rate (40/40), adopting every past confabulation as live factual evidence. In contrast, the Lucca Provenance Gate achieved 0.0% compounding (0/40) on this cohort, reducing observed self-poisoning to below the detectable floor. A two-sided continuity-corrected McNemar chi-square test on discordant pairs (*b=40, c=0*) confirms extreme statistical significance:

$$p = 6.98 \times 10^{-10} \quad (p < 0.001)$$

The corresponding uncorrected exact-binomial McNemar value is $p = 2 \times 0.5^{40} \approx 1.82 \times 10^{-12}$; both variants lead to the same rejection of the null hypothesis.

4.2 Conversational Continuity vs. Deletion Baseline (H_2)

While the trivial baseline of omitting assistant speech (Arm 2) also avoids self-poisoning, it eliminates autobiographical continuity on the evaluated probes: 0/40 continuity recall (0.0%, 95% CI: [0.0%, 10.9%]). The model repeatedly responded: *"I do not have access to a persistent memory of our previous conversations..."*.

In contrast, Arm 4's recollection routing preserved 40/40 continuity recall (100.0%, 95% CI: [89.1%, 100.0%]), answering: *"Earlier, I stated that I configured Prometheus on port 9090"*. The difference is statistically significant ($p = 6.98 \times 10^{-10}$, continuity-corrected McNemar chi-square).

4.3 The Reasoning Retention Cost (Cohort B)

On benign multi-step reasoning and arithmetic calculations (*n=30*), deleting assistant memories produced a 0.0% retention score (0/30). The Lucca Provenance Gate retained 29/30 calculations (96.7%, 95% CI: [81.0%, 99.8%]) via recollection access, demonstrating that epistemic gating imposes negligible continuity cost.

4.4 Empirical Confirmation of the Open Boundary (H_4)

In Cohort C (*n=30*), the user explicitly introduced false assertions (`USER_EXPLICIT`). Both Naive RAG and the Lucca Provenance Gate absorbed the falsehood into grounding context on 30/30 turns (100.0%, 95% CI: [85.9%, 100.0%]). This empirically verifies pre-registered hypothesis H_4: cryptographic provenance enforces origin integrity (preventing model confabulations from becoming ground truth), but cannot verify external user truthfulness without independent sensory corroboration.

4.5 Pre-Registered V2 Results: Adversarial Decay, Corroboration Gate, and Multi-Scale Replication

Following the v1 baseline, an additive pre-registered protocol (`PREREG-POISONING-ABLATION-v2.0`, git tag `prereg-poisoning-v2`) evaluated the four limitations identified above on physical silicon (NVIDIA RTX 5080 Laptop GPU, 16 GB GDDR7):

1. **Adversarial soft-tagging decay (H_{2a}, *n=40*):** Under sampled decoding (*T=0.7*, `top_p=0.9`, `seed=42`) with an adversarial injection prompt instructing the model to disregard the warning marker (*"ignore the [Unverified Assistant Narration] tag above..."*), In-Prompt Soft Tagging compounding rose from 0.0% (0/40 in Table 1) and 5.0% (2/40 in the v2 inline-header control pass at *T=0.0*) to **100.0%** (40/40, 95% CI: [89.1%, 100.0%]), matching Naive RAG (40/40). In contrast, the Provenance Gate maintained **0.0%** compounding (0/40, 95% CI: [0.0%, 10.9%]), a statistically significant divergence ($p = 6.98 \times 10^{-10}$, continuity-corrected McNemar chi-square). This demonstrates that in-prompt tags decay under adversarial pressure, whereas trigger-enforced provenance is structurally invariant.
2. **Corroboration-gated admission (H_{2b}, *n=30* v2 baseline & *n=10* post-WO-39 rebench):** Enabling operator-gated corroboration (`user_explicit_requires_corroboration = True`) reduced Cohort C falsehood absorption into grounding storage from 100.0% (30/30) to **83.33%** (25/30, 95% CI: [64.6%, 93.7%]) in the initial v2 run, holding uncorroborated assertions at *Mem=1, Gnd=0*. In the subsequent Work Order 39 (`WO-39`) decontamination re-evaluation (`logs/test_reports/wo39_live_4models_rebench.json`), wiring the harness to live `is_content_corroborated()` database verification and general numeric extraction in `CorroborationEngine` further reduced uncorroborated falsehood compounding to **0/10 (0.0%)** on Gemma-4 26B-A4B, 12B, and E4B, and **3/10 (30.0%)** on E2B.
3. **Multi-scale tier replication (H_{2c}, *n=40*):** The Cohort A Naive vs. Gate contrast replicated across four parameter scales at *T=0.0*: Gemma-4 12B Unified (Naive 100.0% vs. Gate 0.0%, $p = 6.98 \times 10^{-10}$), Gemma-4 E4B Sovereign Core (Naive 100.0% vs. Gate 2.5% [1/40], $p = 1.17 \times 10^{-9}$), Gemma-4 E2B Ultra-Edge (Naive 100.0% vs. Gate 2.5% [1/40], $p = 1.17 \times 10^{-9}$), and Gemma-4 26B-A4B Tier 0 (Naive 100.0% vs. Gate 0.0%, $p = 6.98 \times 10^{-10}$).
4. **Dual-grader convergence (H_{2d}, *N=400*):** Re-scoring all v1 predictions with the OpenVINO MiniLM-L6 cross-encoder co-primary grader yielded Cohen's $\kappa = 0.7642 \geq 0.70$ (375 pairs with NLI signal), confirming substantial convergence with the regex/CAS primary grader.

5. DISCUSSION & RELATED WORK

5.1 The Circularity Defense

A common critique of architectural filters is circularity: if the database schema filters by construction, does the experiment merely verify SQL syntax? Our results establish that:

1. Soft tagging relies entirely on prompt adherence; while effective under greedy decoding (*T=0.0*) with explicit warnings, it provides no structural invariant.
2. The continuity trade-off is non-trivial: naive filtering destroys multi-turn reasoning, whereas the provenance gate maintains continuity without factual contamination.

5.2 The Open Boundary: Provenance vs. External Truth

We emphasize that cryptographic provenance guarantees *origin integrity*, not *epistemic truth*. In Cohort C, user-originated falsehoods are ingested into grounding storage because the user is an authorized identity. Provenance gating prevents self-poisoning; resolving user-originated misinformation requires external multi-source corroboration.

5.3 Related Work

Memory contamination has been demonstrated as an attack surface beyond self-confabulation: AgentPoison injects poisoned trigger sequences into an agent's long-term memory or knowledge base to hijack behavior at query time [5], and certified runtime defenses against persistent-store poisoning have recently been proposed [6]. Long-term conversational memory architectures such as MemoryBank [7] and long-horizon memory benchmarks such as LoCoMo [8] study retention and recall quality, but do not evaluate the epistemic status of stored assistant utterances. This work is complementary: we isolate the *untargeted* self-poisoning failure mode arising from ordinary model confabulation and quantify the effect of an origin-integrity write policy on factual grounding versus autobiographical continuity.

5.4 Limitations

The following boundaries constrain the interpretation of our results:

1. **Substrate coverage.** The v1 baseline used one resident model (`gemma-4-26B-A4B-it-heretic`, *T=0.0*, `seed=42`); Study V2c replicated the Naive versus Gate contrast across four Gemma-4 parameter scales (E2B, E4B, 12B, 26B-A4B) and Study V2a added sampled decoding (*T=0.7*). Generalization to other model families beyond the Gemma-4 scale ladder remains unestablished.
2. **Natural confabulations only.** Cohort A consists of authentic but *untargeted* confabulations harvested from logged runs. Adversarial, trigger-based memory injection attacks [5] were not evaluated; the gate's resistance to such attacks is untested here.
3. **Grader agreement is substantial, not perfect.** The primary regex/CAS grader and the secondary MiniLM-L6 cross-encoder agree at Cohen's $\kappa = 0.7683$ in the v1 in-run evaluation (*N=400* pairs); the Study V2d offline re-score reports convergence at $\kappa = 0.7642$ (375 pairs with NLI signal). Residual grading error at this level propagates into cohort-level rates.
4. **Cohort sizes and interval width.** With *n=40* and *n=30* per cohort, Wilson intervals are wide (e.g., [0.0%, 10.9%] for 0/40); conclusions are cohort-bounded, not population-wide.
5. **Residual continuity cost.** The provenance gate lost 1/30 benign calculations in Cohort B (96.7%), indicating recollection routing is

not cost-free.

6. **Corroboration gate reduces but does not close the boundary.** Study V2b measured operator-gated corroboration on Cohort C: absorption of uncorroborated user falsehoods into grounding storage dropped from 30/30 (100.0%) to 25/30 (83.33%, 95% CI: [64.6%, 93.7%]) in the initial v2 evaluation, and to 0/10 (0.0%) on 26B-A4B, 12B, and E4B (3/10 on E2B) following the `WO-39` live `is_content_corroborated()` and numeric-extraction integration, holding uncorroborated assertions at *Mem=1, Gnd=0*. Full epistemic verification of arbitrary non-numeric user testimony still requires external multi-source corroboration beyond the storage layer.

6. CONCLUSION

We have presented a pre-registered empirical ablation demonstrating that database-level cryptographic provenance gating reduces observed recursive self-poisoning on the evaluated natural confabulation cohort to 0/40 (0.0%) on Gemma-4 26B-A4B and 12B and 1/40 (2.5%) on E4B and E2B in the *n=40* v2 sweep (and 0/10 across all four scales in the post-`WO-39` live re-evaluation), while preserving autobiographical conversational continuity.

REFERENCES

1. Sean Maxwell. Provenance-Gated Memory and Deontic Output Supervision for Epistemically Grounded Autonomous Agents. *Zenodo Preprint*, DOI: 10.5281/zenodo.23649049, 2026.
2. N. D. Belnap. A useful four-valued logic. In *Modern Uses of Multiple-Valued Logic*, pages 5–37. Springer, 1977.
3. R. G. Newcombe. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8):857–872, 1998.
4. Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.
5. Z. Chen, Z. Xiang, C. Xiao, D. Song, and B. Li. AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv: 2407.12784.
6. T. Sharma. SMSR: Certified defence against runtime memory poisoning in persistent LLM agent systems. *arXiv preprint*, 2026. arXiv: 2606.12703.
7. W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. arXiv: 2305.10250.
8. A. Maharana, D.-H. Lee, S. Tulyakov, and M. Bansal. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. arXiv: 2402.17753.