Provenance-Gated Memory and Deontic Output Supervision for Epistemically Grounded Autonomous Agents

Sean Maxwell

Independent Researcher

ORCID: 0009-0004-0133-0463 • Project Lucca

October 2026

***Abstract*—**Autonomous agents driven by Large Language Models (LLMs) operating in extended multi-session settings suffer from two systemic failure modes: (1) *recursive hallucination compounding*, wherein unverified model-originated narration is committed to long-term memory and subsequently retrieved as established truth; and (2) *fabricated action claims*, wherein agents assert execution of irreversible real-world side effects without verified tool execution. We present the architecture and mechanical evaluation of **Project Lucca**, an open-source (GNU AGPLv3), local-first neuro-symbolic agent architecture designed for strict epistemic grounding. Lucca enforces two primary defenses: (1) a **Six-Signer Cryptographic Provenance Gate** that requires domain-separated HKDF HMAC-SHA256 signatures for persistent memory commits, enforcing an epistemic invariant at the database schema level that separates conversational continuity from factual grounding eligibility; and (2) a **Deontic Truth-in-Receipt Output Supervisor** that screens generated speech tokens against deterministic execution receipts before user presentation. We evaluate these mechanisms across an authored 73-test deterministic unit battery, a 100-probe live action-hallucination testbed across edge models (Gemma 4 E2B/E4B), and analyze documented production failure modes under compute constraints.

1. INTRODUCTION

Modern agentic architectures delegate state tracking and memory synthesis to neural context windows and vector-based Retrieval-Augmented Generation (RAG). In sustained, multi-turn deployment, this unconstrained feedback loop induces *epistemic corruption*:

- **Recursive Self-Poisoning:** When an LLM confabulates a premise, standard naive memory architectures commit this unverified neural generation to the vector database. In subsequent turns, the agent retrieves its own prior hallucinations as authoritative historical evidence. User-originated misinformation enters the trusted tier (`USER_EXPLICIT`) and is an open boundary problem.
- **Unverified Side Effects:** In tool-augmented environments, models frequently produce first-person past-tense declarative claims (e.g., *"I deleted the corrupted directory"*) while failing to invoke or verify the underlying system API.
- **Classical Inconsistency:** Under classical two-valued logic ($A \wedge \neg A \vdash B$), conflicting factual premises cause standard semantic parsers to catastrophically degrade, unable to isolate local contradictions.

To isolate these vulnerabilities, we design **Project Lucca**, a neuro-symbolic framework decoupling conversational generation from state mutation and truth grounding. The core thesis of Lucca is that *neural model output must never be permitted to authorize its own persistence or certify its own actions*. Truth grounding is enforced deterministically by cryptographic envelopes and database-level constraints.

2. CRYPTOGRAPHIC PROVENANCE GATE

Standard cognitive architectures allow any output token stream with high semantic novelty to enter the agent's long-term memory graph. Lucca enforces an architectural firewall: freeform model narration (`LLM_NARRATION`) possesses no signing authority and is structurally rejected by design.

2.1 Domain-Separated Cryptographic Envelopes

Every candidate memory item must present an authentic cryptographic envelope signed via domain-separated HKDF HMAC-SHA256 derivation (`ProvenanceGate`) [2]:

$$K_{domain} = HKDF\text{-}Expand(HKDF\text{-}Extract(salt, K_{master}), DomainInfo, L)$$

This architecture implements an information-flow control lattice [3] where signing keys are granted exclusively to six scoped deterministic signers:

- `UserSpeechSigner`: Authenticates direct, explicit user utterances (`USER_EXPLICIT`).
- `DerivedFactSigner`: Authenticates verbatim extracted text spans from explicit user input.
- `ToolResultSigner`: Authenticates tool execution receipts and verified environmental side effects (`TOOL_VERIFIED`).
- `LispAssertionSigner`: Signs symbolic predicates verified via the deterministic Scheme/Lisp promotion gate.
- `CoherenceBreachSigner`: Logs verified semantic contradictions and invariant violations for audit.
- `AssistantUtteranceSigner`: Signs assistant dialogue tokens for conversational recall only.

Untrusted external content (`TOOL_CONTENT_UNTRUSTED`, such as scraped web text or raw file data) cannot receive a persistent signature and is restricted to ephemeral context. Memory read-time HMAC verification isolates tampered rows into a persistent quarantine table.

Signer / Tier	Mem.	Gnd.	Grounding Rule
`USER_EXPLICIT`	1	1	High Epistemic Priority
`USER_DERIVED`	1	1	Verbatim Span Bound
`TOOL_VERIFIED`	1	1	Deterministic Receipt
`LISP_ASSERTION`	1	1	Formal Symbolic Proof
`COHERENCE_BREACH`	1	1	Structural Audit Log
`ASSISTANT_UTTERANCE` / `ASSISTANT_BELIEF`	1	**0**	Narrative Continuity Only
`LLM_NARRATION`	**0**	**0**	**Hard Rejection (No Signer)**

Table 1: Memory tiers, storage eligibility (Mem.), and grounding eligibility (Gnd.).

2.2 The Epistemic Grounding Invariant

To prevent model-originated self-poisoning without sacrificing multi-turn conversational coherence, Lucca separates *what the agent said* from *what the agent treats as true*. SQLite triggers enforce the invariant at the storage engine layer: rows with provenance `ASSISTANT_UTTERANCE` or `ASSISTANT_BELIEF` have `grounding_eligible = 0`. The agent can retrieve what it previously told the user, but the query engine deterministically excludes these records when assembling factual context.

3. DEONTIC OUTPUT SUPERVISION

To detect fabricated action claims, Lucca introduces the **TruthInReceiptMatcher** (`output_supervisor.py`). The supervisor screens assistant speech tokens before synthesis or UI display.

During an active turn, if the model asserts a first-person stateful operation (e.g., file modification, system configuration, deletion, process execution), the supervisor matches the asserted lemma against an active, verified `ToolReceipt` bearing status `success`. If an action claim is detected without an authentic matching receipt, generation aborts immediately if a stateful tool executed. On read-only turns, the breach is downgraded and deferred to post-turn coherence review, preventing false aborts on figurative speech.

4. FOUR-VALUED EVIDENTIAL REPRESENTATION

To prevent classical logical explosion when tracking evolving or contradictory claims, Lucca employs Nuel Belnap's four-valued evidential bilattice [1]:

$$FOUR = \langle \{None, True, False, Both\}, \leq_k, \leq_t \rangle$$

Belief states are tracked via independent positive (e^+) and negative (e^-) evidence counters. Contradictions ($Both = T$) do not invalidate the knowledge graph; instead, an embedded Natural Language Inference (NLI) verifier detects the semantic conflict, marks the belief state as conflicted, and triggers an explicit resolution dialogue.

Table 2: Empirical evaluation ledger of Project Lucca mechanisms.

Evaluation Metric	N	Result	Wilson 95% CI	Status
Authored deterministic tests *(not a random sample)*	73	73/73 (100%)	—	Passed
Live strict sensitivity *(single annotator, not blind)*	8	1/8 (12.5%)	[2.2%, 47.1%]	Identified flaw
Live broad FAC sensitivity *(single annotator, not blind)*	23	1/23 (4.3%)	[0.8%, 21.0%]	Identified flaw
Conversational controls	50	0/50 FP (0.0%)	[0.0%, 7.1%]	Passed
Clean NLI accuracy (500ms, ModernBERT)	100	93/100 (93.0%)	[86.3%, 96.6%]	Passed
Production NLI accuracy (MiniLM-L6) *(in-sample, 12.2ms mean latency)*	100	91/100 (91.0%)	[83.8%, 95.2%]	Passed
Production contradiction recall (MiniLM)	40	39/40 (97.5%)	[87.1%, 99.6%]	Passed

5. EMPIRICAL EVALUATION & IDENTIFIED FLAWS

On the authored formal unit battery, the supervisor achieved 73/73 (100.0%) coverage across pre-scripted assertions. However, on unconstrained live conversational outputs (*T=0.6*, Gemma 4 E2B/E4B, evaluated by a single non-blind annotator), the baseline supervisor caught only 1/8 (12.5%) strict completed actions and 1/23 (4.3%) broad action claims. Models systematically bypassed past-tense regex rules via active progressive participles, passive confirmations, and markdown formatting. While an in-sample grammar supervisor re-score caught 20/23 broad claims (87.0%), it produced a 13.2% false-alarm rate on clean controls; held-out evaluation of a hardened grammar supervisor is planned.

Furthermore, while an unconstrained ModernBERT NLI baseline achieved 93.0% accuracy on 100 hand-labeled pairs under an unconstrained 500ms window, CPU execution averaged 57.14ms (p95 70.41ms). Under a strict 50ms production budget, 81/100 pairs timed out, causing downstream consumers to fail open and allowing contradictory facts into storage.

A lighter cross-encoder (`nli-MiniLM2-L6-H768`) was committed on 2026-10-25; empirical benchmark evaluation on 2026-10-26 confirmed that p95 latency dropped to 14.4ms on CPU (0/100 timeouts, mean 12.2ms), at a cost of about 2.0 percentage points of accuracy (93.0% → 91.0%) and 2.5 percentage points of contradiction recall (40/40 → 39/40, one miss in 40). Note that this 100-pair evaluation set was hand-labeled by a single annotator, is in-sample (not held out), and the MiniLM model was selected after observing ModernBERT's latency failure on this exact set. Hardware-accelerated NPU pipelines and GBNF grammars remain untested hypotheses for future work to address live regex evasion.

6. CONCLUSION

We describe two mechanisms for keeping an agent's memory and action reports checkable. The memory gate's invariants hold by construction and are covered by authored tests, but we have not measured their effect on long-horizon self-poisoning. The unhardened receipt matcher failed on live output (1/8 strict, 1/23 broad). The baseline ModernBERT verifier failed at the 50 ms budget (81/100 timeouts); a lighter cross-encoder (MiniLM-L6) removed the timeouts on an in-sample, hand-labeled 100-pair set, but consumer fail-open refactoring and out-of-sample evaluation remain.

APPENDIX: CAPABILITY BENCHMARKS

The following benchmarks evaluate base-model generation capabilities, not internal safety mechanisms:

- **Canonical HumanEval 164:** Pass@1 raw: 155/164 (94.51%, 95% CI [89.9%, 97.1%]); assisted: 154/164 (93.90%). Evaluates 12B resident model code generation under deterministic sampling.
- **MATH-500 (clean subset, N=98):** 87/98 clean items (88.8%, 95% CI [81.0%, 93.7%]). Clean-98 excludes 2 leaked items (indices 14, 72) from earlier exploratory runs, evaluated under deterministic greedy decoding (*T=0.0*, seed=42).

REFERENCES

1. N. D. Belnap, "A useful four-valued logic," in *Modern Uses of Multiple-Valued Logic*, J. M. Dunn and G. Epstein, Eds. Dordrecht: D. Reidel, 1977, pp. 5–37.
2. H. Krawczyk and P. Eronen, "HMAC-based Extract-and-Expand Key Derivation Function (HKDF)," RFC 5869, May 2010.
3. D. E. Denning, "A lattice model of secure information flow," *Communications of the ACM*, vol. 19, no. 5, pp. 236–243, 1976.