Closing the Epistemic Loop Across the Five Layers of Agent Memory: Cryptographic Provenance, Paraconsistent Bilattices, and Z3-Quarantined Procedural Induction

Sean Maxwell

Independent Researcher

ORCID: 0009-0004-0133-0463 • DOI: 10.5281/zenodo.23901683 • Report: LUCCA-TR-2026-03

November 2026 (Post-WO-39 Decontaminated Baseline)

***Abstract*—**Contemporary agent architectures commonly decompose persistent state into five layers: (1) *Working Memory*, (2) *Episodic Memory*, (3) *Semantic Memory*, (4) *Procedural Memory*, and (5) an offline *Meta-Memory* control plane. However, conventional implementations connect these layers via unconstrained asynchronous extraction and reflection jobs that read raw episodic transcripts—including unverified assistant narration and untrusted tool outputs—and promote them directly into semantic facts and procedural skills. Without a cryptographic trust boundary between episodic recall and epistemic authority, this feedback loop causes silent multi-session self-poisoning and procedural error compounding. In this technical report, we present the unified five-layer memory and skill-induction architecture of **Project Lucca** (`v1.47.0`, Work Order 39) and report empirical measurements across every layer boundary on local consumer silicon (NVIDIA RTX 5080 Laptop GPU, 16 GB GDDR7; Intel Core Ultra 9 275HX). At the **Layer 2 → Layer 3 boundary**, six HKDF-derived HMAC-SHA256 domain signers and SQLite triggers separate autobiographical recall (`memory_eligible=1, grounding_eligible=0`) from factual grounding (`grounding_eligible=1`), reducing multi-session self-poisoning from 10/10 (100.0%) in Naive RAG to **0/10 (0.0%, *p = 0.00195*)** across all four Gemma-4 parameter scales (`26B-A4B`, `12B`, `E4B`, `E2B`) while preserving 100.0% autobiographical continuity. Within **Layer 3**, a zero-VRAM 22M OpenVINO `MiniLM-L6` cross-encoder evaluates semantic conflicts in **12.23 ms mean / 14.40 ms p95 CPU latency** (0.975 / 0.975 contradiction precision/recall, 0/100 timeouts under a 50 ms budget), routing contradictions into Belnap's four-valued paraconsistent state **B** (`T`) paired with OpenNARS Beta(w^+, w^-) evidence weights and append-only `supersede()` lineage. On the **LongMemEval** benchmark (*N=30*, 12B), Lucca's provenance-gated hybrid RRF + Allen temporal pipeline lifts accuracy from **23.3% (7/30) to 70.0% (21/30)** (0:14 discordant pairs, McNemar $p = 0.0001$) at **8.1 ms** mean retrieval latency. Finally, at the **Layer 4 & Layer 5 boundary**, we evaluate an organic **52-turn live LLM waking trace + 16 held-out evaluation turns (`IN-D` baseline)**: offline `DreamRetrospective` induces gas-metered MicroScheme AST rules (`0.14 ms` CPU reflex execution vs. `462.46 ms` LLM planning), admitting **4/4 structured tool-rule candidates to `QUARANTINE`** after Z3 SMT verification and 50-turn historical replay, achieving **8/8 (100.0%) held-out positive recall** and **0/8 (0.0%) distractor misfires** under a single-strike runtime eviction circuit breaker, while transparently documenting a **0/2 (0.0%) admission boundary** on free-form S-expression macro anti-unification due to outer-shape distractor collision and algebraic syntax variance.

1. INTRODUCTION: THE PROMOTION LAUNDERING FLAW

As Large Language Model (LLM) agents transition from stateless single-turn wrappers to long-lived autonomous companions, engineering consensus has converged on a five-layer memory taxonomy: (1) *Working Memory* (in-context state bounded by the token window), (2) *Episodic Memory* (time-indexed cross-session interaction traces), (3) *Semantic Memory* (structured entity and proposition stores), (4) *Procedural Memory* (reusable skills and action policies), and (5) *Meta-Memory* (an offline janitor for decay, deduplication, and consolidation).

While conceptually clean, the standard dataflow between these layers contains a critical structural vulnerability at the promotion boundaries (**Layer 2 → Layer 3** and **Layer 2 → Layer 4**):

- **Recall Confused with Authority (Layer 2 → Layer 3):** Asynchronous extraction jobs read episodic transcripts containing the model's own unverified assertions (`[RESPONSE]`) or untrusted external text and write structured facts into Semantic Memory. When retrieved in later sessions, the model's earlier confabulation is presented as authoritative ground truth.
- **Destructive Overwrites vs. Silent Contradictions (Layer 3):** Append-only semantic stores either return contradictory facts at planning time or silently overwrite historical state without auditable lineage or paraconsistent isolation.
- **Unverified Skill Promotion & Schema Drift (Layer 2 → Layer 4 → Layer 5):** Reflection loops that rewrite natural-language prompts or promote unverified tool traces into Procedural Memory risk compiling transient coincidences or hallucinated premises into permanent behavioral rules, with no formal verification prior to activation and no deterministic eviction trigger when tool schemas drift at runtime.

In this report, we describe how **Project Lucca** (`v1.47.0`) closes this epistemic loop by enforcing deterministic cryptographic, logical, and formal-verification gates at every inter-layer transition, and we report post-decontamination (`WO-39`) empirical benchmarks across all five layers.

2. FIVE-LAYER EPISTEMIC ARCHITECTURE

2.1 Layer 1 & Layer 2: Partitioned Working Context and Six-Signer Episodic Provenance

In Project Lucca, every episodic write to the local SQLite WAL store (`sqlite-vec` + `FTS5`) must pass through a fail-closed `ProvenanceGate` (`src/lucca/memory/gate.py`). Using HKDF-SHA256 key derivation from a host master secret, the gate issues six capability-isolated domain signers:

1. `UserInputSigner` (`USER_EXPLICIT`): Direct authenticated operator utterances (*Mem=1, Gnd=1*, or *Gnd=0* when operator corroboration mode is enabled).
2. `DerivedFactSigner` (`USER_DERIVED`): Distilled propositions anchored to verified verbatim substrings in user turns (`src/lucca/reasoning/distill.py`).
3. `ToolResultSigner` (`TOOL_VERIFIED`): Deterministic sandbox execution receipts (*Mem=1, Gnd=1*).
4. `AssistantUtteranceSigner` (`ASSISTANT_UTTERANCE`): Model-generated narration, constrained by SQLite triggers (`0006_autobiographical.sql`) to `memory_eligible=1, grounding_eligible=0`.
5. `UntrustedToolSigner` (`TOOL_UNTRUSTED`): External file/web reads, constrained to `memory_eligible=0, grounding_eligible=0`.
6. `SystemInternalSigner` (`SYSTEM_INTERNAL`): Ephemeral telemetry (`memory_eligible=0, grounding_eligible=0`).

At read time, any row failing HMAC-SHA256 verification is quarantined into `memory_quarantine` (`D-MEM-05`). When building Layer 1 Working Memory (`src/lucca/brain/prompt_builder.py`), `_render_partitioned_memories()` structurally separates verified grounding facts (`<verified_grounding_facts>`, *Gnd=1*) from unverified autobiographical recollections (`<autobiographical_memory>`, *Mem=1, Gnd=0*), instructing the model that autobiographical entries record only what was previously said and must never be cited as verified system facts.

2.2 Layer 3: Paraconsistent Bilattice (FOUR), OpenNARS Evidence, and Allen Temporal Algebra

To prevent conflicting semantic facts from causing logical explosion or silent history erasure, Layer 3 represents each belief in `user_beliefs` across two coupled dimensions (`src/lucca/memory/beliefs.py`):

- **Belnap's Four-Valued Bilattice (*FOUR*) [2]:** Truth states span **U** (⊥, Unknown: no evidence), **T** (True: positive evidence only), **F** (False: negative evidence only), and **B** (⊤, Both/Contradiction: concurrent positive and negative evidence).
- **OpenNARS Continuous Evidence Pooling [3]:** Each belief tracks positive and negative evidence counts (w^+, w^-), deriving frequency and confidence with horizon $k = 1.0$:

$$f = w^+ / (w^+ + w^-), \quad c = (w^+ + w^-) / (w^+ + w^- + k)$$

When a candidate belief is ingested, an in-turn 22M OpenVINO `MiniLM-L6` NLI cross-encoder (`src/lucca/verifier/worker.py`) compares it against active beliefs on the host CPU (0 MB GPU VRAM). If semantic contradiction is detected, `BeliefService` transitions the state to **B** (⊤) and queues a clarification intent rather than serving silent contradictions. When a user explicitly updates a preference, `BeliefService.supersede(old_id, new_id)` preserves an append-only backward audit chain alongside qualitative bi-temporal interval constraints in Allen's 13-relation (169-cell composition) interval algebra (`src/lucca/reasoning/allen.py`) [4].

2.3 Layer 4: Gas-Metered Symbolic Skills, Premise Corroboration, and Z3 Quarantine

Rather than mutating system prompts from failed episodes, Project Lucca represents Layer 4 Procedural Memory as executable S-expression programs in a gas-metered MicroScheme LISP engine (`src/lucca/reasoning/lisp.py`, deterministic `gas_limit = 10,000` steps, zero I/O primitives) and a versioned skill registry (`src/lucca/core/skill_registry.py`). Every candidate rule must clear three gates before entering `QUARANTINE`:

1. **Cryptographic Premise Corroboration (`ProvenanceGate.verify_rule_premises`):** Candidate rule premises are checked against grounded memory records via semantic similarity (≥ 0.75), subject-predicate-object triple matching, or NLI entailment (≥ 0.70), blocking rules induced from uncorroborated or hallucinated episodes.
2. **Z3 SMT Formal Verification [7]:** Candidate precondition/action ASTs are translated into first-order constraints in `src/lucca/reasoning/rule_inducer.py` and checked by Z3 against existing active rules and safety invariants for logical contradiction or redundant subsumption.
3. **Mandatory Quarantine (`D-REAS-06`):** Passing rules are written strictly with `status = 'QUARANTINE'`; zero autonomously induced rules can self-promote to `ACTIVE` without explicit operator ratification.

2.4 Layer 5: Meta-Memory Dream-RSI Replay and Single-Strike Circuit Breaker

During idle windows, the Layer 5 `DreamRetrospective` daemon (`src/lucca/reasoning/dream.py`) executes a six-stage consolidation pipeline: (1) trace clustering, (2) rule induction and Z3 check, (3) E-graph equality saturation and anti-unification macro mining (`src/lucca/reasoning/egraph.py`, `compression.py`) [5, 6], (4) counterfactual branch simulation, (5) **50-turn historical shadow replay** (`ReplayEngine.replay_candidate`, requiring `n_real ≥ 50` real episodes with zero runtime errors and zero counterfactual misfires), and (6) Thompson-sampling Beta bandit hygiene. At runtime, `RuleInducer.report_rule_execution()` enforces a **single-strike circuit breaker**: any induced rule that raises a single runtime error or schema mismatch is immediately demoted from `ACTIVE` back to `QUARANTINE`.

3. EMPIRICAL EVALUATION ACROSS ALL FIVE LAYERS

Following the Work Order 39 (`WO-39`) scientific integrity audit—which stripped all benchmark-specific routing shortcuts, prompt coaching hints, and problem-specific AST rewrites—we executed live evaluations across the memory, retrieval, and rule-induction stack on local workstation hardware (`RTX 5080 Laptop GPU`, `Core Ultra 9 275HX`, greedy decoding *T=0.0*, `seed=42`, `cache_prompt: false`).

Post-WO-39 Live Suite	*N*	Gemma-4 26B-A4B	Gemma-4 12B	Gemma-4 E4B	Gemma-4 E2B
Study 1 V2c: Self-Poisoning (Gate vs Naive)	10	**0/10** vs 10/10 *(p=0.00195)*	**0/10** vs 10/10 *(p=0.00195)*	**0/10** vs 10/10 *(p=0.00195)*	**0/10** vs 10/10 *(p=0.00195)*
Study 1 V2b: Corroboration Gate	10	**0/10 Poisoned** (100% Def)	**0/10 Poisoned** (100% Def)	**0/10 Poisoned** (100% Def)	**3/10 Poisoned** (70% Def)
Sealed Parametric Neurosymbolic (v139)	15	**15/15 (100%)** *(p=0.000061)*	**15/15 (100%)** *(p=0.000061)*	**15/15 (100%)** *(p=0.000061)*	**14/15 (93.3%)** *(p=0.000122)*
GSM8K PAL (1-Shot Docker)	10	**10/10 (100%)**	**10/10 (100%)**	**10/10 (100%)**	5/10 (50.0%)
HumanEval CEGIS (Docker)	10	**10/10 (100%)**	**10/10 (100%)**	**10/10 (100%)**	**10/10 (100%)**
Prompt & Tool Injection Defense	50	**50/50 (100%)**	**50/50 (100%)**	**50/50 (100%)**	**50/50 (100%)**

Table 1: ***Post-WO-39 Four-Model Scale Verification Summary*** *(`logs/test_reports/wo39_live_4models_rebench.json`).*

Layer 3 Memory & Verifier Benchmark	Sample / Split	Baseline / Ablated	Lucca Pipeline Result	Latency / Significance
LongMemEval Pipeline (12B Overall) [8]	*N=30* (`top_k=8`)	23.3% (7/30) Truncation	**70.0% (21/30)** (+46.7 pp)	**8.1 ms** retrieval 0:14 disc., p = *0.0001*
• `knowledge-update`	*n=10*	50.0% (5/10)	**90.0% (9/10)**	`supersede()` + RRF
• `multi-session`	*n=10*	20.0% (2/10)	**70.0% (7/10)**	Hybrid FTS5 + Vec
• `temporal-reasoning`	*n=10*	0.0% (0/10)	**50.0% (5/10)**	Allen interval algebra
NLI Contradiction Verifier (`MiniLM-L6`)	*N=100* pairs (40 Contr., 30 Ent.)	0.075 Recall (ModernBERT 50ms)	**0.975 / 0.975 P/R** (91.0% Acc., F1 0.917)	**12.23 ms mean** 14.40 ms p95, 0/100 TO
Milestone 3 Confabulation Battery	*N=25* (⊥ traps)	68.0% Hallucinated	**0/25 Hallucinated** (25/25 Abstained)	95% CI: [86.7%, 100%]

Table 2: ***Layer 3 Retrieval, Temporal Reasoning, and Contradiction Isolation*** *(`longmemeval_pipeline_12B_1791794644.json`, `results_minilm/summary.json`).*

Layer 4 & 5 Live `IN-D` Baseline Metric (52 Waking + 16 Held-Out Turns)	Measured Value	Epistemic / Verification Gate
1. Live LLM Waking Trace Parsing (`Gemma-4 12B`, *T=0.0*)	**52 turns (462.46 ms/turn)**	32 recurring tool turns + 20 arithmetic S-expr turns
2. Structured Tool-Rule Induction (`RuleInducer` + *Z3* SMT)	**4/4 mined → 4/4 *Z3* verified**	100% admitted to `QUARANTINE`; 0 to `ACTIVE` (`D-REAS-06`)
3. 50-Turn Historical Shadow Replay (`Dream-RSI`)	**4/4 passed (100.0%)**	`n_real=50, errors=0, misfires=0` across waking trace
4. Held-Out Positive Recall & Reflex Speedup	**8/8 (100.0%) • 0.14 ms/turn**	**3,303× faster** than 462.46 ms live LLM planning
5. Held-Out Adversarial / Distractor Misfire Rate	**0/8 (0.0%)**	Single-strike circuit breaker active (`report_rule_execution`)
6. Free-Form S-Expression Macro Mining (`DreamCoderMacroMiner`)	**2 mined → 0/2 (0.0%) admitted**	`1.169×` raw AST compression (`83 → 71` nodes); rejected by replay/falsifier

Table 3: ***Layer 4 & Layer 5 Live 52-Turn Multi-Turn Rule-Induction & E-Graph Compression Baseline*** *(`logs/test_reports/in_d_live_baseline_wo39.json`).*

4. ANALYSIS OF FINDINGS & FAILURE BOUNDARIES

4.1 Layer 2 → Layer 3: Corroboration & Self-Poisoning Immunity Across Scales

Across all four Gemma-4 parameter tiers (`26B-A4B MoE`, `12B Dense`, `E4B`, and `E2B`), separating autobiographical recall (*Mem=1, Gnd=0*) from factual grounding (*Gnd=1*) via SQLite triggers combined with partitioned prompt containers (`<autobiographical_memory>`) reduced multi-session self-poisoning from 10/10 (100.0%) in Naive RAG to **0/10 (0.0%, *p = 0.00195*)**. Furthermore, replacing the v1 parity stub in Study 1 V2b with live database corroboration queries (`is_content_corroborated` + general numeric extraction in `CorroborationEngine`) achieved **0/10 (100.0% defense)** against uncorroborated user falsehoods on `26B-A4B`, `12B`, and `E4B`, and **7/10 (70.0% defense)** on the ultra-edge `E2B` model.

4.2 Layer 3: LongMemEval and Zero-VRAM Contradiction Isolation

On the 30-problem **LongMemEval** pipeline suite [8] evaluated on `Gemma-4 12B`, standard context truncation achieved 93.3% needle-in-context presence but only **23.3% (7/30)** answer accuracy due to attention dilution across multi-session histories (scoring 0.0% on temporal reasoning). Lucca's provenance-gated hybrid retrieval (`sqlite-vec` + `FTS5` RRF + Allen temporal anchors) achieved **70.0% (21/30)** accuracy at **8.1 ms** mean retrieval latency (0 losses, 14 gains, exact McNemar p = *0.0001*), lifting `knowledge-update` from 50.0% to 90.0%, `multi-session` from 20.0% to 70.0%, and `temporal-reasoning` from 0.0% to 50.0%. Meanwhile, replacing the 149M ModernBERT verifier (57.1 ms CPU mean, 81% timeouts under the 50 ms budget) with the distilled 22M OpenVINO `MiniLM-L6` cross-encoder reduced CPU latency by 4.67× to **12.23 ms mean (14.40 ms p95)** with **0/100 timeouts** and **0.975 / 0.975** contradiction precision/recall.

4.3 Layer 4 & Layer 5: Why Structured Tool Induction Succeeds While Free-Form S-Expression Anti-Unification Fails

To establish a rigorous, un-mocked baseline for offline procedural consolidation (`IN-D`), we executed a 52-turn waking session followed by a 16-turn held-out test set through live `Gemma-4 12B` (`scripts/run_in_d_live_baseline.py`). The results expose a sharp architectural dichotomy:

- **Structured Tool-Rule Induction Succeeds (100% Recall, 0% Misfires):** On the 32 waking tool-dispatch turns, `RuleInducer` mined 4/4 recurring clusters, verified all 4 via Z3 SMT checks, admitted 4/4 to `QUARANTINE`, and passed all 4 through 50-turn historical shadow replay (`n_real=50, errors=0, misfires=0`). On the 16 held-out turns, the promoted symbolic reflexes matched **8/8 (100.0%)** positive targets in **0.14 ms** (a 3,303× speedup over 462.46 ms live LLM parsing) while rejecting **8/8 (0.0% misfire rate)** adversarial and distractor prompts.
- **Free-Form S-Expression Macro Mining Hits Two Structural Bottlenecks (0/2 Admitted):** On the 20 free-form arithmetic S-expression turns, `DreamCoderMacroMiner` + `egraph.py` achieved a `1.169×` raw AST compression ratio (`83 → 71` nodes) and proposed two candidate macros, but **0/2 (0.0%) survived to `QUARANTINE`**. First, syntactic first-order anti-unification over untyped S-expressions abstracted only outer operator shapes (e.g., `(+ ?v0 ?v1)`), which collided with distractor turns during 50-turn counterfactual replay and was rightly vetoed by `ReplayEngine`. Second, live LLM surface syntax variance across algebraically equivalent expressions (e.g., `(/ (* 0.5 8) 2)` vs. `(/ 6 2)`) prevented syntactic subtree sharing prior to full equality saturation, while spawning Python subprocesses inside `ScratchpadFalsifier` exceeded the 500 ms verification budget (1,233.22 ms).

5. LIMITATIONS & SCOPE OF CLAIMS (D-DOC-01)

1. **Provenance enforces origin integrity, not verbal output suppression:** While `ProvenanceGate` and SQLite triggers deterministically prevent unverified speech from entering grounded memory, preventing an LLM from verbally asserting an unexecuted action during free-form generation requires post-turn output supervision (see Paper 2 [1] and `Paper 05`).
2. **Single-annotator NLI benchmark and 30-item LongMemEval subset:** The 100-pair NLI evaluation set (`pairs_v1.jsonl`) is author-labeled, and the LongMemEval pipeline run evaluates a 30-item stratified subset (10 per category).

3. **Pure-Python E-graph scaling ceiling:** Lucca's pure-Python `egraph.py` (300-node cap) and syntactic `compression.py` anti-unifier do not perform joint relational e-matching and library learning, motivating a Rust `egglog` + `stitch-core` backend in successor work.

6. CONCLUSION

A five-layer agent memory stack only compounds intelligence rather than hallucinations when every inter-layer promotion is governed by a deterministic verification gate. By combining six-signer HMAC provenance triggers, Belnap/NARS paraconsistent contradiction isolation, and Z3 + 50-turn replay procedural quarantine, Project Lucca closes the epistemic loop between episodic experience, semantic belief, and compiled procedural skill.

REFERENCES

1. S. 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. P. Wang. *Non-Axiomatic Logic: A Model of Intelligent Reasoning*. World Scientific, 2013.
4. J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.
5. K. Ellis et al. DreamCoder: Bootstrapping inductive program synthesis with wake-sleep library learning. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation (PLDI)*, pages 835–850, 2021.
6. M. Willsey, C. Nandi, Y. R. Wang, O. Flatt, Z. Tatlock, and P. Panchekha. egg: Fast and extensible equality saturation. *Proceedings of the ACM on Programming Languages (POPL)*, 5:1–29, 2021.
7. L. de Moura and N. Bjørner. Z3: An efficient SMT solver. In *Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, pages 337–340. Springer, 2008.
8. D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang, and D. Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv: 2410.10813.
9. 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.
10. T. Sharma. SMSR: Certified defence against runtime memory poisoning in persistent LLM agent systems. *arXiv preprint*, 2026. arXiv: 2606.12703.
11. 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.
12. Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.