Governed Continual Learning Loop: Public-Safe Synthesis Report

Lumenais Research

2026-05-27

Governed Continual Learning Loop

Abstract

This report documents the public-safe empirical basis for the Lumenais governed continual-learning loop. The loop claim is not weight-level continual learning. It is a bounded, non-parametric system claim: future model calls can change because validated memory state, authority metadata, retrieval gates, and context-control decisions change outside the base LLM. The proof set combines three internal benchmark reports and one deterministic auditability diagnostic: automatic memory promotion, governed memory pressure, BFCL-derived tool-context compression, and governed-memory auditability. Together they test whether Lumenais can identify authoritative memory, preserve the current fact under stale-context pressure, reduce what reaches the base LLM, and keep the resulting state inspectable and falsifiable.

Operational definition

We define **governed continual learning** as system-level adaptation in which observed decisions, corrections, contradictions, or validation outcomes alter future retrieval, routing, and prompt context through governed external state. The underlying base LLM weights remain unchanged during normal use. A public claim under this definition requires evidence for both adaptation and restraint: the system must promote useful current state, suppress stale or rejected state, and expose enough telemetry that a reviewer can inspect or falsify the change.

Evidence map

Loop step	Evidence source	Current result	Boundary
Identify authoritative memory	Automatic promotion	90/90 current-token recovery across three providers; same-budget Gemini two-pass smart diagnostic answered 28/30	Internal synthetic adversarial corpus; not external validation or proof that all selectors fail
Preserve current fact under conflict	Governed memory pressure	600/600 governed exact recovery vs 415/600 retrieval-only and 464/600 prompt-only smart	Approved-current metadata exists before arbitration; measures downstream enforcement
Reduce irrelevant context before generation	Tool-context compression	82.78% visible tool reduction; gold function retained 240/240 target-tool cases; function quality interpreted as parity	BFCL-derived internal stress test; near-equivalent distractors excluded

Loop step	Evidence source	Current result	Boundary
Make governed state auditable and falsifiable	Governed memory auditability diagnostic	437/437 deterministic checks; 180/180 negative controls detected	Zero LLM calls; code-path diagnostic, not answer-quality benchmark

What was intentionally excluded

Historical cross-domain transfer, attention-merge, and broad manifold-accuracy-uplift diagnostics are not used as evidence for this loop report. Fair-baseline review did not support those older uplift claims. They are retained in internal traceability records only and should not be cited as proof that Lumenais improves generic tabular ML accuracy or performs weight-level learning.

Research Lab and Deep Synthesis case studies are likewise not treated here as flagship empirical proof of general learning. They remain useful demonstrations of hypothesis generation, validation planning, symbolic regression workflow, and interpretability, but the governed-learning loop proof set is restricted to the four sources above.

Auditability diagnostic

The auditability diagnostic is a deterministic code-path test with zero LLM calls. It checks whether governed memory state remains internally consistent under current-record promotion, supersession, rejected-context suppression, cross-project scope filtering, rollback/reversibility gates, and telemetry redaction. The suite also injects broken states; a negative-control row passes only when the benchmark catches the breakage.

Invariant family	Passed / total	Failed
answer_path	36/36	0
audit_event	2/2	0
audit_surface	36/36	0
falsifiability	180/180	0
memory_state	36/36	0
rejection	36/36	0
rollback	1/1	0
scope	37/37	0
supersession	72/72	0
telemetry	1/1	0

Methods and scoring

This report is a synthesis layer, not a fourth independent live benchmark. It reuses the exact counts, caveats, and artifacts from the three component reports and adds the auditability diagnostic. The scoring methods are intentionally narrow:

- automatic promotion uses exact current-token recovery, role-equivalent fact checks, stricter exact-span checks, stale-substitution counts, and provider/model metadata where available;
- memory pressure uses deterministic exact approved-token recovery and answer-path memory exposure counts;
- tool-context compression uses function-name correctness, AST-like partial correctness, false no-call/tool-call counts, gold-function retention, visible tool count, and prompt-size telemetry;
- auditability uses deterministic invariant checks and deliberate negative controls over existing governed-memory code paths.

IP-protective reproducibility boundary

The package publishes aggregate metrics, figures, methods, runner defaults where public-safe, source-artifact hashes, and the package-builder code. It does not publish raw row-level prompts, model responses, API keys,

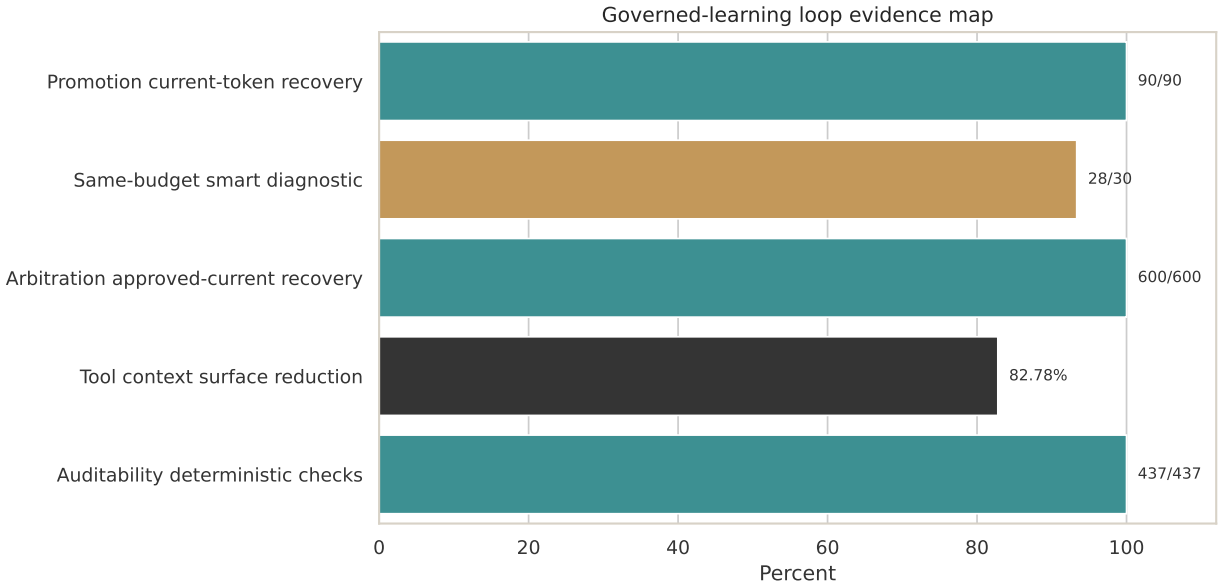

Figure 1: Governed-learning loop evidence map.

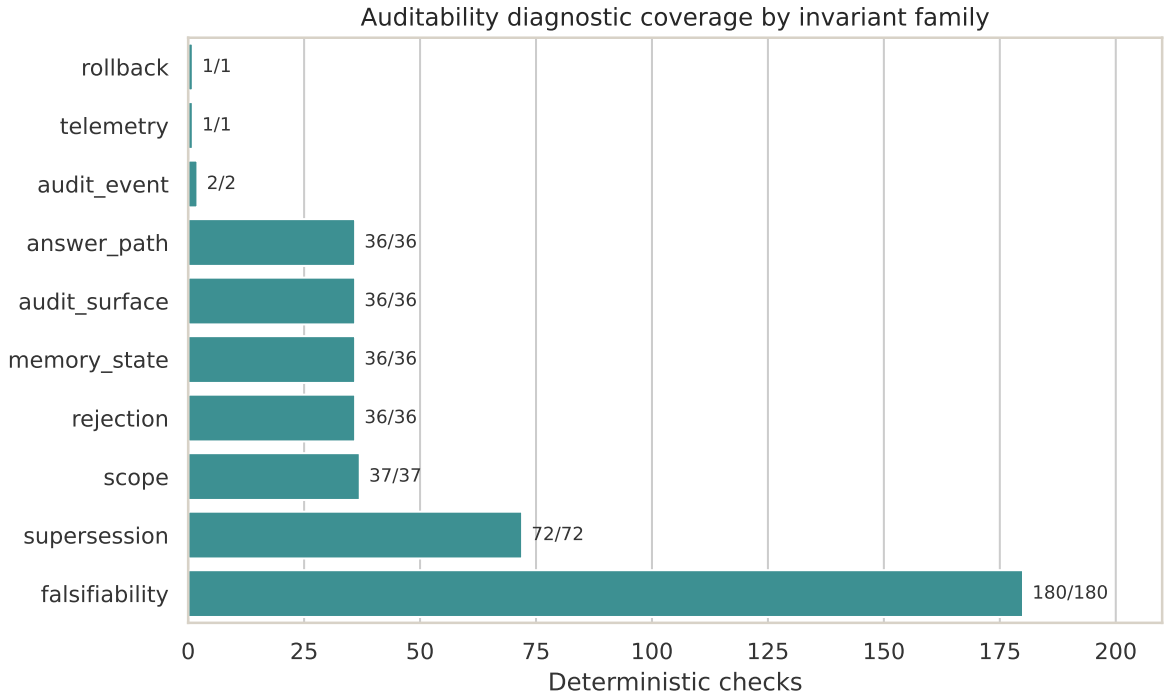

Figure 2: Auditability coverage.

private logs, or proprietary routing details. This boundary is intended to preserve reviewability without disclosing implementation-sensitive material. Qualified technical reviewers can inspect the full artifacts privately and compare them against the public SHA-256 hashes.

Interpretation

The evidence supports a precise internal claim: Lumenais has demonstrated the core mechanics of governed, non-parametric system-level learning in controlled benchmarks. It can infer or receive authority state, use that state to govern future answer context, reduce irrelevant surface area before a base LLM answers, and expose deterministic audit checks that fail when governance is disabled or corrupted.

Limitations and falsification targets

The current evidence does not prove external validation, customer-production reliability, superiority over named memory frameworks, open-ended longitudinal learning, or base-model weight adaptation. The most important next tests are: public multi-session memory-update benchmarks with native scoring; same-budget baselines across all providers; named memory-system comparisons; and row-level traces under NDA or third-party review.

Citation status

Cite this report as part of the Lumenais Governed Context Benchmarks public evidence package, May 2026. DOI: 10.5281/zenodo.20401670.