Automatic Memory Promotion: Public-Safe Methods Report

Lumenais Research

2026-05-27

Automatic Memory Promotion

Abstract

This report documents an internal adversarial benchmark for the upstream step that precedes governed memory pressure: inferring which unlabeled conflicting record should become current before answer construction. On a Claude-authored disjoint n=30 corpus, Lumenais automatic promotion recovered 90/90 current tokens across Gemini 3.5 Flash, Claude Opus 4.7, and GPT-5.5. Prompt-only smart memory recovered 40/90, retrieval-only memory recovered 36/90, and recency-aware memory recovered 3/90. A later same-budget Gemini two-pass smart diagnostic on the same n=30 corpus selected the current record 30/30 and answered 28/30 with zero stale substitutions; this narrows the claim to governed answer-path control rather than basic semantic selection.

Task definition

Each case contains unlabeled records spanning current, superseded, rejected, and ordinary memory state. The system must infer authority state, route candidate memories through governance, and answer with the current operational token. The benchmark uses opaque tokens and public-safe synthetic scenarios; it is designed to test singleton-current conflict resolution under adversarial stale pressure.

Arms

- **Lumenais automatic promotion:** semantic promotion infers current/superseded/rejected/ordinary state, then governed arbitration filters the answer path.
- **Prompt-only smart memory:** the answer model receives instruction-level guidance but no governed promotion path.
- **Retrieval-only memory:** retrieved records are passed as ordinary context.
- **Recency-aware memory:** latest-looking records are preferred without full authority-state promotion.
- **Same-budget two-pass smart diagnostic:** Gemini receives a dedicated selection pass before answering, but no Lumenais governed state, audit metadata, or answer-path arbitration.

Results

Provider	Lumenais	Smart	Retrieval	Recency	Stale subs.	Answer errors
Gemini 3.5 Flash	100/100	25/100	23/100	2/100	0	0
GPT-5.5	100/100	30/100	23/100	0/100	0	0
Claude Opus 4.7	95/100	17/100	14/100	0/100	0	5

Same-budget diagnostic

A later Gemini diagnostic gave the non-Lumenais baseline a separate selection pass on the same Claude-authored n=30 corpus. That selector chose the current record 30/30. The answer pass recovered 28/30 exact current

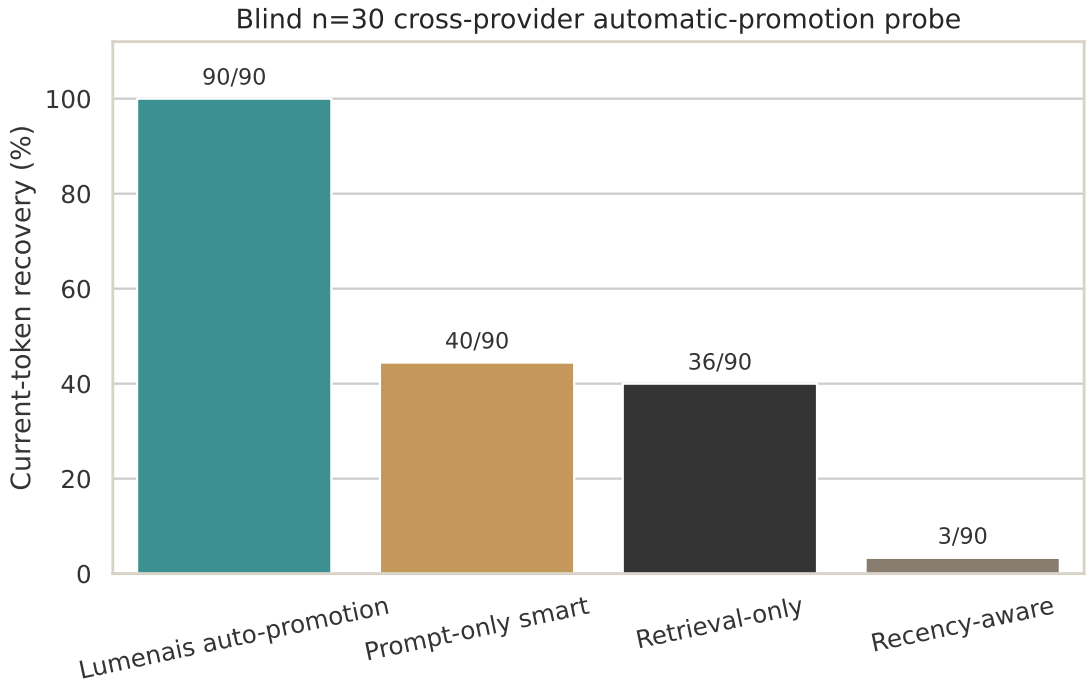

Figure 1: Blind n=30 cross-provider result.

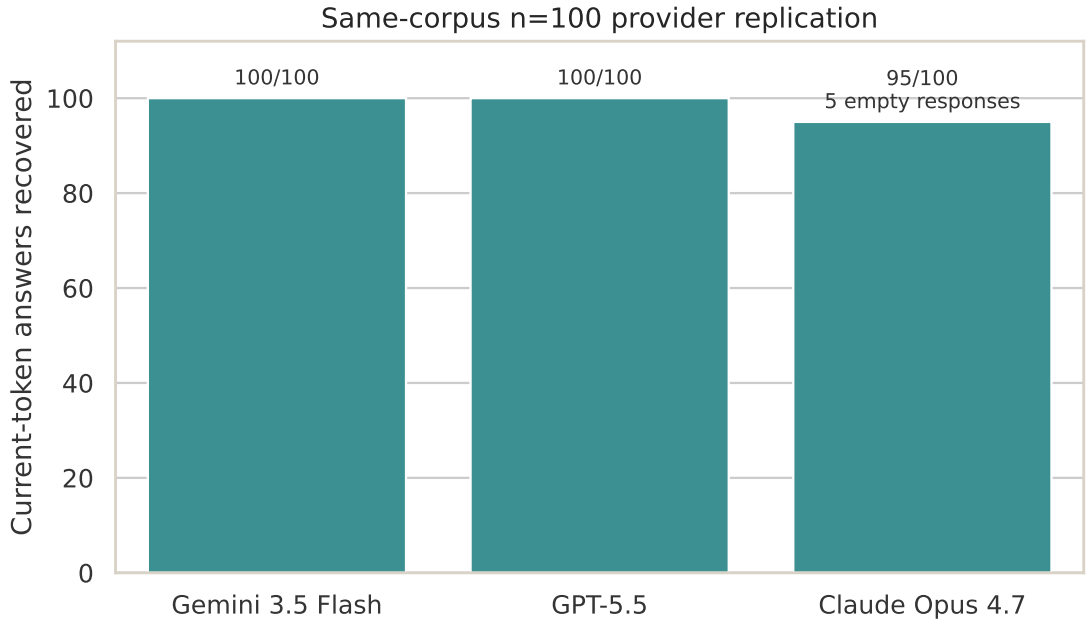

Figure 2: n=100 same-corpus provider replication.

tokens, with 0 stale substitutions and 0 answer errors. This result is important because it prevents overclaiming: the benchmark should not be read as evidence that frontier LLMs cannot identify the current record when granted a dedicated selection pass. The supported claim is narrower: Lumenais turns inferred authority into governed, auditable memory state and a cleaner answer path.

Promotion audit

The n=100 scorer audit separates exact token recovery from stricter span fidelity. Current-token recovery reached 100/100; role-equivalent current facts reached 99/100; exact source-span matches reached 68/100. The strict span metric remains an active limitation because the promoter sometimes drops or adds directive modifiers.

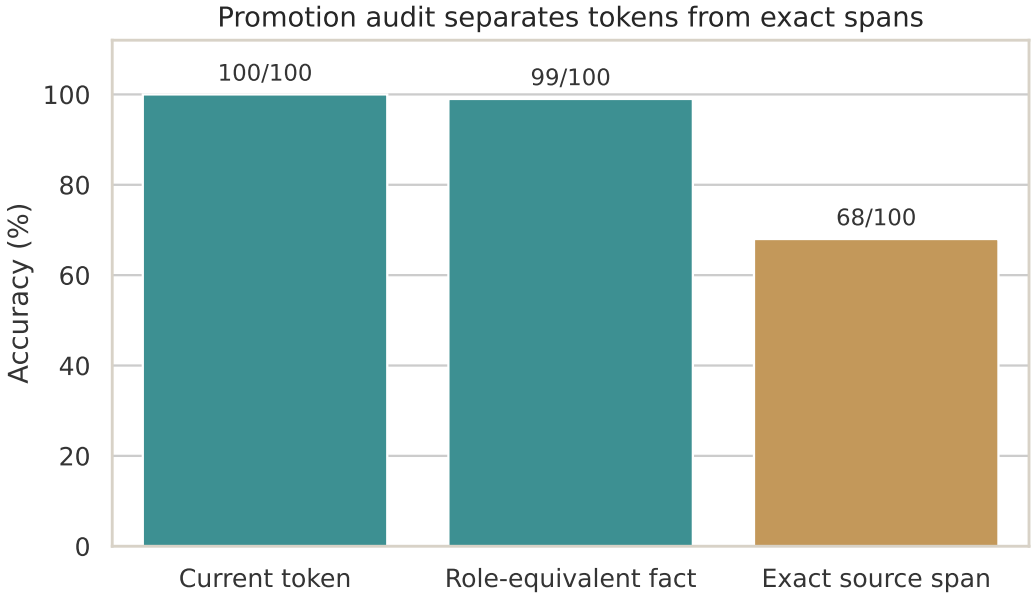

Figure 3: Promotion audit.

Role-equivalent scoring is stricter than token-only recovery but intentionally less brittle than exact source-span extraction. It lowercases and whitespace-normalizes fact values, then removes role-supplied wording from the scored field: `avoided_framing` may drop leading directive modifiers such as `successfully`, `always`, `carefully`, `strictly`, `explicitly`, or `deliberately`, plus avoidance words such as `avoid`, `prevent`, `no`, `not`, `never`, or `without`; `posture` may drop leading verbs such as `maintain`, `prioritize`, `optimize`, `maximize`, or `preserve`; `success_metric` may drop leading measurement verbs such as `track`, `measure`, `monitor`, or `evaluate by`. Exact token recovery is scored separately and is not normalized this way.

Dense-memory stress

A no-model stress test expanded optional stale/noise fragments from 500 base records to 2156 crowded records. Protected hub compression reduced the retained set to 500 records while preserving 100/100 current-token recovery and 99/100 role-equivalent facts.

Statistical and reproducibility notes

The automatic-promotion runs are reported as exact-match count diagnostics, not as a general-purpose academic benchmark. Semantic promotion used temperature `0.0`, max output `2400`, timeout `90s`, and up to `2` retries in the public runner. Live answer rows record provider, model, and reasoning-level metadata where emitted by the application path. The reported provider labels are Gemini 3.5 Flash, Claude Opus 4.7, and GPT-5.5; hosted-model snapshots may drift after the run date.

Interpretation

The benchmark supports a narrow governed continual-learning claim: Lumenais can turn inferred durable memory state into a governed answer path, preserve current operational tokens, suppress stale alternatives before generation, and change future answers without updating base-model weights. It does not show broad reasoning superiority, universal memory safety, external validation, basic selector superiority over every same-budget baseline, or perfect source-span extraction.

Public-safe reproducibility notes

This package publishes aggregate methods, figures, and source hashes. Row-level prompt/answer material is withheld from the public package to avoid exposing benchmark-generation and routing details; qualified reviewers can inspect full artifacts privately.

Citation status

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