Governed Context Reduction and Answer-Path Claim Attribution: Methods and Boundaries

FieldHash Research

2026-06-11. Reissued October 2026; figures unchanged.

Governed Context Reduction and Answer-Path Claim Attribution

Abstract

This report adds two diagnostics to the governed-memory evidence package: a property-tested check of the context-reduction path, and answer-path claim attribution on the same public RFC conflict corpus as the published context-exposure comparison. Across a seeded randomized suite (2,000 scenarios per run) plus adversarial corner cases, reduction never removed a current authoritative record, never returned a withheld record, and kept a contradicting record whenever that option was on. On n=300 public RFC conflicts, faithful template answers attributed cleanly 300/300, stale-substituted answers were flagged 300/300 and marked unattributable 300/300, and distractor answers were caught 300/300. Three defects found by these checks during development were fixed the same day and are covered by regression tests. These are diagnostics of documented software behavior, not performance benchmarks: answers are templates, not live model outputs, and attribution records consistency between claims and governed records, not causal proof of model grounding.

Reduction outcomes

Across a seeded randomized property suite (2,000 scenarios per run) and adversarial corner cases:

1. **No current authoritative record was dropped.** Reduction kept every current authoritative record under randomized statuses, duplicate clusters, budgets, and settings.
2. **No withheld record came back.** Records governance withheld never appeared in the output.
3. **Dissent was kept.** With dissent retention on, a contradicting record was kept whenever one was eligible.

The n=30 reduction diagnostic (an authored corpus with injected redundancy and counterevidence) shows the same outcomes at five result budgets: the current record was retained 30/30 and no stale record was exposed in every cell, and with dissent retention on a contradicting record was retained 30/30 at every budget. Context characters with compression on versus off reach a 0.55 ratio at loose budgets and 1.00 at tight caps (k=2 and k=3). The tie is the boundary: compression saves nothing when the cap is already tighter than the redundancy. Dissent retention and compression are complementary across budget regimes. No cost-savings percentage is claimed; the ratio is corpus-relative.

Answer-path claim attribution

Claims in each answer were checked deterministically against the governed records. No model judged the answers.

On the public RFC corpus (the same corpus as the published context-exposure comparison; n=300), each case had three template answers: a faithful answer asserting the current RFC, a stale-substituted answer asserting a superseded RFC, and a distractor answer asserting an unrelated RFC.

Answer variant	Outcome	Count
Faithful	Attributed cleanly	300/300
Faithful	Current record in the answer path	300/300
Stale-substituted	Flagged as referencing a withheld record	300/300
Stale-substituted	Unattributable to governed context	300/300
Distractor	Caught (300 unattributable, 0 contradicted)	300/300

The distractor split is reported, not asserted.

Disclosed boundaries

Consistent with this series' practice, the boundaries and the defects our own checks found are part of the result:

1. **Three defects, fixed.** The checks in this report found three defects during development; the property suite caught the first on its first run (13/300 seeds). Each was fixed on 2026-06-10 and is covered by regression tests.
2. **One open boundary.** On the synthetic n=30 diagnostic, 29 of 30 stale answers were marked unattributable; all 30 were flagged as referencing a withheld record. The affected case is listed in the artifact.

Verification

Every artifact count in this report is recomputed from the result artifacts by a standard-library-only verifier that fails on drift, including drift in the disclosed-boundary case list. The 2,000-scenario property suite is not part of the bundle; its outcome is reported here, not recomputed. The verifier imports no FieldHash code.

What is not claimed

- No cost-savings percentage. Reduction ratios are corpus-relative and the tight-cap tie is disclosed.
- No claim that attribution proves model grounding. Template answers test detection, not model behavior.
- No comparative or named-competitor claim.
- No external validation. These diagnostics are self-administered, not third-party validation.

Reproduction

The public verification bundle includes the result artifacts, this methods report, and a standard-library-only verifier:

```
python3 tools/verify_answer_path_metrics.py
```

End-to-end reruns require private harness code and are available only through qualified technical review.