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

FieldHash Research

2026-06-11

Governed Context Reduction and Answer-Path Claim Attribution

Abstract

This report adds two mechanism diagnostics to the governed-memory evidence package: a property-tested contract for the context-reduction path, and answer-path claim attribution evaluated on the same public RFC conflict corpus as the published context-exposure comparison. The reduction contract is verified by a seeded randomized suite (2,000 scenarios per run) plus adversarial corner cases: no sequence of compression, gating, or capping operations removes a protected record; withheld records cannot re-enter the answer path; and when the counterevidence floor is enabled, at least one eligible contradicting record survives reduction with the conflict reported rather than smoothed over. The attribution diagnostic builds claim packets from the governed answer path for three template answer variants per case. On n=300 public RFC conflicts, faithful answers attributed cleanly 300/300, stale-substituted answers were flagged 300/300 (the asserted identifier traces to a withheld record, and the claim is unattributable to governed context), and distractor answers were caught 300/300. Three defects found by these instruments during development were fixed the same day, regression-pinned, and are disclosed below. These are mechanism claims about documented software behavior, not performance benchmarks: answers are templates, not live model outputs, and attribution records consistency relations with disclosed methods, not causal proof of model grounding.

The reduction contract

Context reduction in FieldHash is the composition of a status gate (rejected, superseded, and ignored records are withheld), non-protected context compaction, a Bayesian relevance gate over non-protected rows, and a result cap that is soft against protected rows. The claims:

1. **Compaction cannot drop canonical authority.** Protected records (reviewed-current, or promoted by a governed learning decision) survive every reduction operation. Verified by invariant I1 of the property suite under randomized statuses, duplicate clusters, caps, and toggle combinations.
2. **Withheld records cannot re-enter.** Rejected, superseded, and ignored records never appear in the output, including as compression representatives. Protection never outranks withholding under any toggle combination (invariant I2).
3. **Reduction preserves dissent.** With the counterevidence floor enabled, at least one eligible contradicting record survives reduction. Governance outranks retention: a superseded contradicting record stays withheld (invariant I3).

Measured reduction (internal synthetic corpus with injected redundancy, stated in the artifact config): 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. The floor and compression are complementary across budget regimes. No cost-savings percentage is claimed; the ratio is corpus-relative.

Answer-path claim attribution

The claim-attribution packet extends the evidence packet from record selection to answer-path relations. Claims in an answer are segmented deterministically with character offsets, then mapped to governed records with deterministic methods only: declared-value matching (word-boundary, never substring) and lexical overlap with disclosed thresholds. Conflicts dominate aggregation. Two signals matter most:

- **The leak signal.** If a claim asserts a value that traces to a record governance withheld, the claim is flagged `references_withheld_record`. This signal is value-based and collision-proof.
- **Counterevidence conflicts surface.** If a retained contradicting record conflicts with a claim, the claim's status is `contradicted`; the reduction contract guarantees the dissenting record was available, and the packet guarantees the conflict is reported.

On the public RFC corpus (same index, chains, seed, and distractor construction as the published context-exposure comparison; n=300): faithful template answers attributed cleanly 300/300 with the current record in the answer path 300/300; stale-substituted answers were leak-flagged 300/300 and unattributable 300/300; distractor answers were caught 300/300, all via unattributability, because distractors did not survive into the roughly one-record answer path and the contradiction route therefore never engaged. The split is reported, not asserted.

Disclosed boundaries and instrument-found defects

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

1. **Protection-ordering edge (fixed 2026-06-10).** The property suite's first run (13/300 seeds) found that with the status gate disabled, promoted-then-superseded records became protected, inverting authority ordering at the cap. No production path was affected (the gate is hardcoded on). Fixed so protection requires passing the status check; regression-pinned.
2. **Scope-blind contradiction (fixed 2026-06-10).** The synthetic answer-path diagnostic's first run showed neighbor-scope authority falsely contradicting faithful answers. Contradiction is now scope-gated: a record cannot contradict claims outside its own scope, and a scope match replaces the fuzzy topic gate. Regression-pinned.
3. **Substring value matching (fixed 2026-06-10).** The RFC port's first run showed the raw-substring fallback matching an RFC identifier inside a longer project key, which made the governing record falsely support every answer and blocked contradiction entirely. Now word-boundary matching; regression-pinned.
4. **Lexical collision (open, disclosed).** On the synthetic corpus, one case in thirty shows a token collision earning weak lexical-overlap support that masks the unattributable flag; the value-based leak signal still fires. Value-based methods are collision-proof; token-overlap methods are not. The affected case ids are recorded in the artifact's `disclosed_boundaries`.

Verification

Every figure in these diagnostics is recomputed from the artifact's own row-level data by a standard-library-only verifier that fails on drift, including drift in the disclosed-boundary case lists: the admissions cannot drop without trace between regenerations. The verifier cannot import the code it audits; a test enforces its import whitelist.

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 demonstrate the detection mechanism, not model behavior.
- No comparative or named-competitor claim.
- No external validation. Public-data diagnostics here are self-administered, not third-party validation.

Reproduction

The public verification bundle includes result artifacts, this methods report, and a standard-library-only verifier. It recomputes the page-visible figures without importing FieldHash implementation code:

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

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