Public Authority-Signal Diagnostics: Methods and Boundaries

FieldHash Research

2026-06-09 (v1.1, recovery diagnostic added same day). Reissued October 2026; figures unchanged.

Public Authority-Signal Diagnostics

Abstract

This report adds three public-data diagnostics to the governed-memory evidence package: an RFC supersession run with explicit authority metadata visible, the same RFC conflicts with that metadata hidden from FieldHash, and a Coffee StackExchange messy-corpus diagnostic using accepted answers as a proxy label. The most important result is the negative ablation. With RFC Obsoletes/Obsoleted-By metadata visible, FieldHash enforced the current RFC in 300/300 conflicts and suppressed stale answer-path exposure to 0/300. With the same authority metadata hidden, FieldHash matched semantic retrieval at 74.00% and stale answer-path exposure returned to 285/300. The supported claim is therefore bounded: FieldHash enforces configured authority evidence deterministically; it does not claim autonomous authority inference from arbitrary prose. A v1.1 follow-up diagnostic (n=600) tests the propose-then-approve recovery path on that boundary: proposed supersession links drafted from record prose and approved before enforcement closed the hidden-authority gap fully where relations are stated in field-style prose (600/600, deterministic rules proposer) and 59% where only narrative prose remains (535/600, reviewed LLM proposer).

Task definitions

RFC visible-authority diagnostic. The evaluator builds conflict-rich candidate sets from the public RFC index. Ground truth comes from RFC Editor `Obsoletes` and `Obsoleted-By` metadata. FieldHash receives the authority metadata and governs which RFC record can reach the answer path.

RFC hidden-authority ablation. The same conflicts and same ground truth are used, but explicit authority metadata is hidden from FieldHash and retained only for scoring. This tests whether the current implementation can recover authority from content, dates, and text alone.

Coffee StackExchange messy-corpus diagnostic. The evaluator uses public Coffee StackExchange posts. Accepted answer status is treated as a proxy authority label, not certified truth. This run is useful for do-no-harm and context-reduction signals, not as a flagship supersession proof.

Propose-then-approve recovery diagnostic (v1.1). Using the same hidden-authority configuration at n=600 (hidden baseline 441/600 against a 600/600 visible ceiling), a proposal pass reads each record's text only (no timestamps, review fields, chain membership, or ground-truth flags) and proposes supersession links. Approval is simulated at two bounds: oracle (only proposals matching ground truth are approved; perfect-reviewer upper bound) and auto (all proposals applied; no-reviewer floor). Approved links become governed state and enforcement runs unchanged. Two text conditions are tested: full prose (field-style relation statements present) and stripped prose (field-style statements removed; only narrative mentions remain, simulating corpora without structured supersession fields).

Results

RFC run	n	FieldHash	Baseline	FH-only	Base-only
Visible authority	300	100.00%	74.00%	78	0
Hidden authority	300	74.00%	74.00%	0	0

RFC run	Stale answer-path exposure	Context reduction
Visible authority	0/300	84.7%
Hidden authority	285/300	66.5%

Coffee proxy result. On the Coffee StackExchange semantic score10 run, FieldHash scored 64.67% versus 54.00% for the baseline across n=300 rows, a net gain of +32. The strongest safety finding was 0/288 no-conflict false demotions with 64.0% context reduction.

The RFC visible-authority result is highly significant against semantic retrieval by exact McNemar test (FieldHash-only wins 78, retrieval-only wins 0, p = 6.617e-24). The hidden-authority ablation is intentionally negative (p = 1.0); hiding the metadata collapses the governed path to retrieval-like behavior.

Recovery diagnostic results (n=600, gap = 159 conflicts).

Arm	Governed correct	Gap closed	Stale-exposure rows	Link precision / recall
Hidden baseline (no proposals)	441/600	n/a	563	n/a
Field-style prose, rules proposer (oracle or auto)	600/600	100%	0	0.99 / 1.00
Narrative only, rules proposer	511/600	44%	310	0.98 / 0.33
Narrative only, LLM proposer, reviewed	535/600	59%	195	0.88 / 0.51
Narrative only, LLM proposer, unreviewed	525/600	53%	196	0.88 / 0.51

All recovery arms are significant against the hidden baseline (field-style rules: +159/-0, p = 2.7e-48; reviewed LLM: +107/-13, p = 1.5e-19). The LLM proposer beats the same-budget narrative rules proposer (+31/-7, p = 1.2e-04). Review decided the outcome for LLM proposals: skipping it cost ten net selections (+11/-1, p = 0.0063) and doubled regressions from 13 to 23.

Disclosed failure mode. When review approved only part of a supersession chain, the wrong record could be selected. Thirteen conflicts regressed even under perfect review. Pilots should approve whole chains before

enforcement.

Proposer-model robustness. The recall ceiling is set by the text, not the model: only 30-40% of gold links have their counterpart mentioned in the stripped text at all, and the budget proposer (Gemini 3.1 Flash Lite) captured 99% of the mentioned signal. A stronger proposer (Gemini 3.5 Flash) scored within 0.4pp on link precision and slightly lower on recall (527/600 reviewed, +0/-8 paired against the budget model).

Interpretation

The public-data evidence now supports two separable claims:

- **Enforcement with authority evidence:** when authority signals are available, FieldHash can translate them into deterministic answer-path control and keep obsolete alternatives out of the model packet.
- **No autonomous prose-inference claim:** when those authority signals are removed, the current governed path does not reliably infer supersession from RFC prose alone.

This is the commercially relevant boundary for pilots. The first scoping question should be where authority signals already live: contract lifecycle systems, policy-management tools, SharePoint/document version history, approval workflows, review states, amendment links, effective dates, rollback events, or systems of record. FieldHash is positioned to enforce those signals on the AI answer path and leave verifier-ready evidence. If those signals are absent, the appropriate workflow is propose-then-approve metadata creation before deterministic enforcement, not live autonomous gating. The v1.1 recovery diagnostic quantifies that workflow's bounds: complete recovery where supersession is stated in structured prose, partial recovery (59% of the gap) where only narrative remains, with human review measurably decisive and the partial-chain failure mode disclosed.

Artifacts

Row-level result artifacts ship in the verification bundle (`data/raw/`); `CHECKSUMS.txt` pins every member.

Limitations

These diagnostics are not third-party validation, customer-production validation, or a named-competitor comparison. RFC metadata is cleaner than many enterprise corpora. Coffee StackExchange accepted answers are proxy labels and can diverge from recency or community consensus. The report should be read as evidence for deterministic enforcement of visible authority signals and for a falsified boundary on arbitrary-prose authority inference. In the recovery diagnostic, oracle approval is a perfect-reviewer upper bound and auto approval a no-reviewer floor; a real review queue lands between them.

Citation status

Cite this report as part of the FieldHash Governed Memory Benchmarks public evidence package.