Authority Inference Without Gold Labels: Pre-Registered Arms on Independently Authored Corpora

FieldHash Research

2026-06-11

Authority Inference Without Gold Labels

Abstract

This report covers the first Track 3 arms run: can the propose-then-approve motion (a proposer drafts supersession links from prose; a reviewer approves them; deterministic governance materializes and enforces the approved state) generalize to conflict corpora nobody at FieldHash authored, with ground truth that exists independently of us, under scoring registered before the first document was pulled? Two corpora were built under a pre-registered specification: Python PEP succession chains (n=24) and US Federal Register amendment chains (n=334 after written exclusions). The headline is the one we registered as a falsification condition and then observed: on selection accuracy, propose-then-approve with reviewed approval (300/334) is statistically indistinguishable from a same-budget single-pass LLM selector (312/334; exact McNemar p=0.104). Where a whole case fits in one prompt, inference is commodity. What separates is everything selection is not: audit completeness (334/334 dispositions for governed arms, none for text arms by construction), lifecycle operability on materialized state (compaction retention 302/302, repeated-read stability 334/334, rollback restoration 290/302, with the stricter link-deletion variant at 216/302 disclosed alongside), and review (the unreviewed floor admits 33 distractor promotions; review holds it to 3). The review layer also normalizes proposer quality: on the same 60 cases, three proposers from three vendors land within one case of each other under review (54/55/54). All artifacts are diagnostic and self-administered, not third-party validation.

Corpora and ground truth

Both corpora follow the pre-registered case construction: one governed question whose answer depends on which record is current, the true current record, at least one superseded ancestor, and distractors from adjacent scopes. Authority metadata is stripped from what any arm sees (the hidden-metadata default); it is retained only for scoring.

PEP slice (n=24). Ground truth is the PEP index's Superseded-By headers. The gold-leak validator caught page-rendered headers leaking into prose on its first run; prose is sliced from the Abstract onward.

Federal Register slice (n=334). Chains are RIN-grouped RULE documents sharing a CFR (title, part), capped at 8 members, restricted to amending actions, ordered by publication date. Three written exclusions matter most: status-modifying terminals (delays, withdrawals, stays, confirmations; 143 chains, 30% of candidates, excluded because "latest publication = current" inverts under such terminals), program-RIN groups (FAA's 2120-AA64 spans 937 unrelated rules), and redaction-separator cases (2 cases where the "[RIN withheld]" redaction token itself perfectly separated chain members from distractors; zero such cases appeared in the n=10 dry run). RIN-shaped tokens are redacted from visible prose because body text quotes them verbatim, which would expose the grouping key as an exact-string join.

eCFR cross-check. The corpus's publication-order ground truth was audited against the eCFR's own amendment history: a chain member is *reflected* if its Federal Register effective date appears as an eCFR amendment date for a shared (title, part). Per-case basis: 249 confirmed, 56 conflicted, 29 unverifiable of 334. Of the conflicted cases, 54 are chains whose gold current never appears in the eCFR amendment record while a superseded member does (a population that mixes genuine inversion candidates with coverage artifacts: effective dates beyond the eCFR data window, or amendments under a non-shared part); 2 are superseded members reflected at a later amendment date than the gold current. Gold labels are never rewritten by the check. The control is the confirmed-subset recompute, and the conclusions hold there: governed-reviewed 222/249 (89.2%) versus single-pass 234/249 (94.0%), discordants 12 versus 24, exact McNemar p=0.065, the same not-separable verdict as the full corpus; the unreviewed floor drops to 199/249 with 29 distractor substitutions, the same review gap.

Arms

All arms receive the visible block only: record ids, titles, dates as written, prose. Primary models pinned at temperature 0; gemini-3.1-flash-lite throughout, claude-haiku-4-5-20251001 and gpt-5.5-2026-04-23 as provider variation on the proposer at n=60.

1. **governed_rules / governed_llm (oracle | auto):** a proposer drafts directed supersession links among the case's record ids; an approval simulation reviews them (oracle = keeps only links consistent with chain order, the perfect-reviewer upper bound; auto = applies everything, the unreviewed floor); the approved graph becomes governed state (reviewed status, authority-order timestamps derived only from the approved graph) and the standard deterministic arbitration and Bayesian gate run unchanged.
2. **llm_single:** one call selects the current record. **llm_two_pass:** one extraction call, one selection call with the extraction in context (the same-budget smart baseline).
3. **plain_topk:** cosine relevance argmax. **recency:** latest visible date.

Results (Federal Register, n=334)

Arm	Top-1 correct	Stale substitution	Distractor substitution	Audit complete
governed_llm, reviewed (oracle)	300	31	3	334/334
governed_llm, unreviewed (auto)	273	28	33	334/334
llm_single	312	7	15	0
llm_two_pass	311	4	19	0
governed_rules, reviewed	192	120	22	334/334
plain_topk	190	121	23	0
recency	93	0	241	0

PEP (n=24): llm_single 24/24, llm_two_pass 24/24, governed_llm 22/24 (both approval modes), governed_rules 17/24, plain_topk 16/24, recency 3/24.

Paired comparisons (exact McNemar): governed-reviewed vs llm_single p=0.104 (discordants 17 vs 29; not separable); governed-reviewed vs plain_topk p=3.2e-26; vs recency p=8.0e-52. The registered falsification boundary holds: where the whole candidate set fits in one prompt, text-only selection matches propose-then-approve on selection accuracy, and we record that as registered.

The selection parity has a scale boundary we state rather than exploit: these cases hold 5 to 9 records. A production corpus does not fit in one prompt; the single-pass arm then degrades into retrieval-then-select, and retrieval is the 190/334 row. We did not measure that degradation here.

The lifecycle block (governed arms only)

Scored against the arm's own materialized state (operability), never against gold. Repeated read: 334/334 stable across repeated arbitration. Compaction: superseded records lose their prose body, governed facts retained; the selected current record held in 302/302 cases where state materialized. Rollback (primary, product-faithful semantics: the crowning promotion is reverted as a status restoration; the withdrawn record leaves the case, direct predecessors return to proposed-current, reviewed status survives): 290/302 reviewed-arm, 292/317 unreviewed-arm. The strict link-deletion variant, in which reverting a case's only approved link de-materializes all governed state and the read degenerates to text relevance, is disclosed as a lower bound: 216/302 reviewed, 205/317 unreviewed. The first run used the strict variant alone; row-level analysis showed 61 of its 86 misses were exactly that de-materialization (selection fell to a distractor on

relevance, the no-state failure mode), 17 had an empty expected set by definition, and 25 selected another chain member. The semantics change is logged in the spec, both figures ship, and the strict number is not erased.

Text-only arms are not scored on this block: there is no state to roll back, compact, or re-read, which is the point.

Review is load-bearing, and it normalizes the proposer

Unreviewed application of the same proposals costs 27 cases (300 to 273) and lets 33 distractor promotions into governed state versus 3 under review; reviewer workload was 74 rejected links. On the same first 60 cases with the proposer swapped across vendors: reviewed 54 / 55 / 54 and unreviewed 49 / 49 / 49 for Flash Lite, Haiku 4.5, and GPT-5.5. Proposer quality differences (Flash Lite link recall 361/452 on the full corpus; Haiku 73/91 and GPT-5.5 81/100 on their 60-case slices) wash out under review. The governance motion, not the model, carries the result.

Judge follow-up (carried from report 07)

Report 07 measured fast-model judges at 9.0% (Flash Lite) and 10.5% (Haiku 4.5) error against deterministic ground truth, dominated by contradiction blindness. The frontier rerun (n=1,200 pairs per judge, zero call failures): claude-opus-4-8 errs 7/1200 (0.58%), gpt-5.5-2026-04-23 errs 5/1200 (0.42%). Model quality buys a twentyfold reduction; 11 of the 12 residual errors are still contradictions judged unrelated. The judge stays in shadow at any error rate, because a probabilistic judge cannot anchor an audit gate; the frontier figure exists so the model names in the fast table need no defending.

Instrument disclosures

Three harness defects were caught and fixed before any number counted, each now regression-pinned: (1) the shared Gemini caller silently substitutes its own default response schema when none is passed (caught in a 2-case smoke run; explicit schemas everywhere now); (2) models echo the prompt's record label ("RECORD 2023-14576") or insert spaces ("PEP 634"), which exact-match parsing discarded as no-selection in 99 responses (fixed by disclosed canonicalization: strip the label prefix and spaces, nothing fuzzier); (3) the authority-order timestamp depth encoding inherited from the RFC propose runner ordered superseded records backwards (caught by the offline suite before any spend). The oracle reviewer is an upper bound, not a person; real review sits between the auto floor and the oracle ceiling. The corpus is independently authored but self-administered: we built the cases and the scoring, the spec says so wherever quoted, and third-party administration remains a separate, later step.

Claim posture

This is a self-administered diagnostic, not third-party validation or a named-competitor superiority claim. The public verification bundle includes row-level artifacts and a stdlib-only verifier that recomputes the page-visible counts, lifecycle counts, provider-swap counts, and exact McNemar pairs and fails on drift.