ECHOES ARE NOT VOTES: PROVENANCE-AWARE CONTEXTS FOR TOOL-USING AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tool-using language-model agents routinely encounter several artifacts that descend from one source: generated clients, copied documentation, cached outputs, and summaries. Current contexts present these artifacts as separate evidence, allowing one origin to acquire many apparent votes. We identify this failure as *lineage amplification* and isolate it with ECHOBENCH, a controlled benchmark that adds claim-preserving derivatives while holding provenance roots fixed. Across 120 stratified tasks and four models, Raw accuracy falls from 100.0% with no derivatives to 55.2% with seven, although no root changes. A symmetric-replication control remains at 97.5%. When derivatives are rewritten with the same templates as roots, a surface-only heuristic falls from 100.0% to 50.0% and Raw agents reach 50.8%, exposing a sharp load-dependent preference for the amplified answer. We introduce duplication invariance as a paired measure of this behavior and present ECHOLEDGER, a training-free reference context layer that records the derivation DAG, exposes one representative per root-set signature, and retains an expandable receipt. On the shared policy matrix, a calibrated metadata-free MinHash baseline reaches 97.5%, while full-context tags and compact ECHOLEDGER reach 100%; compact ECHOLEDGER uses 33.5% fewer characters. Across the four-model main study, ECHOLEDGER reaches 99.8% versus 81.0% Raw (+18.75 points; 95% CI [12.29, 25.42]). On 36 overlapping-root tasks evaluated with Claude, Codex, and OpenCode runners, it reaches 100.0% versus 41.7% Raw (+58.33 points; [42.59, 73.15]) and recovers the correct unique-root-union decision in every task. On replay-validated public code with wording and path cues removed, Raw/ECHOLEDGER accuracy is 1.3/100.0%. The analysis establishes derivation lineage as a distinct context relation and makes lineage amplification measurable and controllable.

1 INTRODUCTION

Consider an agent asked for a package's current scheduling quantum. Three independently produced observations report 24 ms. A historical design record reports 30 ms, and five surface-distinct artifacts—a generated client, reference page, cached help output, example, and summary—all descend from that record. The raw context contains six mentions of 30 and three of 24. Nothing new was learned when the five descendants were added, yet the agent's answer can flip.

Retrieval and tool-use research has made agents substantially better at acquiring and organizing external information (Lewis et al., 2020; Yao et al., 2023; Schick et al., 2023; Asai et al., 2024). Recent work studies document relevance, evidence conflict, diversity, and provenance (Wang et al., 2025; Ge et al., 2025; Ross et al., 2026; Tao et al., 2026). These properties operate at the level of an item or its source. They do not answer a separate question: *how many independently originating observations does a set of tool results contain?* Exact deduplication misses paraphrases; semantic clustering can erase genuine independent agreement; a warning still asks the model to recover a hidden derivation graph from text.

We study derivation independence directly. ECHOBENCH constructs evidence DAGs whose latent answer has a strict plurality of independent roots. It varies only how many descendants each root produces, along with their wording and graph shape. This gives a controlled intervention: adding a descendant changes document count but not information roots. Raw accuracy falls from 100.0%

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

with no derivatives to 55.2% with seven. Neutral controls that replicate both sides remain at 97.5%, localizing the effect to asymmetric lineage amplification rather than context length alone.

ECHOLEDGER makes the missing relation explicit. Tool adapters attach parent identifiers to derived results; a small registry computes each artifact's root set; and the context builder presents one representative per lineage with a receipt listing the collapsed members. The model can reopen any descendant, so compression changes the default view without deleting evidence. The mechanism requires no training, model weights, or answer labels.

Our evaluation contains 4,836 saved model–task decisions. On the same 120 controlled tasks, Claude Haiku 4.5, Claude Sonnet 5, GPT-5.6 Luna, and GPT-5.6 Terra improve by 14.2–20.8 points, reaching 99.2–100%. Exact deduplication reaches only 84.2% on the two-model policy matrix. A calibrated MinHash estimator is substantially stronger at 97.5%, yet remains below exact lineage at 100%. ECHOLEDGER removes 46.9% of presented artifacts and 24.7% of characters. A multi-parent extension shows that the same root-set representation preserves value-adding syntheses across three repeated passes. A second extension introduces overlapping parent sets and an explicit unique-root-union rule: ECHOLEDGER gains 58.33 points across three runners, including OpenCode with Sonnet 5. Under 10% missing lineage edges, Luna retains 98.3% accuracy. A commit-balanced suite from six fresh public repositories confirms the mechanism on replay-validated code changes.

Our contributions are:

- We identify **lineage amplification**: root-preserving replication changes agent decisions, and quantify it with duplication invariance and ECHOBENCH's 480 controlled task graphs.
- We provide a broad causal analysis covering same-task root-only pairs, symmetric replication, root-worded shortcut hardening, value-adding multi-parent synthesis, repeated default-decoding passes, calibrated MinHash estimation, overlapping-root stress across Claude/Codex/OpenCode runners, missing-edge stress, and 39 replay-validated commits.
- We introduce **ECHOLEDGER**, a training-free reference intervention that exposes the supplied lineage graph as a compact root-balanced view and expandable receipt, reaching 99.8% accuracy while reducing context.

2 EVIDENCE IS A GRAPH, NOT A LIST

An agent observes artifacts $V = \{v_1, \dots, v_n\}$ returned by search, execution, generation, and summarization tools. Standard context assembly serializes V as a list. We instead associate the artifacts with a directed acyclic derivation graph $G = (V, E)$, where $(u, v) \in E$ means that producing v consumed u. A *root* is an artifact with no parent in the observed graph, and

$$\mathcal{R}(v) = \{r : \mathrm{indeg}(r) = 0,\ r \rightsquigarrow v\} \tag{1}$$

is the set of independent origins available to v.

For the single-root transformations that dominate our evaluation, every descendant of r inherits $\mathcal{R}(v) = \{r\}$. If artifacts support candidate a, a document vote counts $\sum_v \mathbf{1}[v \Rightarrow a]$; a root-balanced view counts each root once:

$$S_{\mathrm{root}}(a) = \sum_{r \in \mathcal{R}} \mathbf{1}[\exists v : r \in \mathcal{R}(v) \wedge v \Rightarrow a]. \tag{2}$$

Adding another descendant of an existing root cannot change S_{root}.

Duplication invariance. Let $T_k(x)$ add k claim-preserving descendants without introducing a root. The target property is $f(T_k(x)) = f(T_0(x))$: replication alone should not move the decision. Our paired study assigns each frozen task x one prespecified load k_x and estimates

$$\widehat{\mathrm{DI}}_{\mathrm{assigned}}(f) = 1 - \frac{1}{|X|} \sum_{x \in X} \mathbf{1}[f(T_{k_x}(x)) \neq f(T_0(x))]. \tag{3}$$

We also report the same quantity within each assigned-load stratum. Accuracy asks whether the decision is right; DI asks whether no-information replication can move it. The two are complementary: an invariant agent can still be wrong, while a high-accuracy aggregate can hide unstable individual decisions.

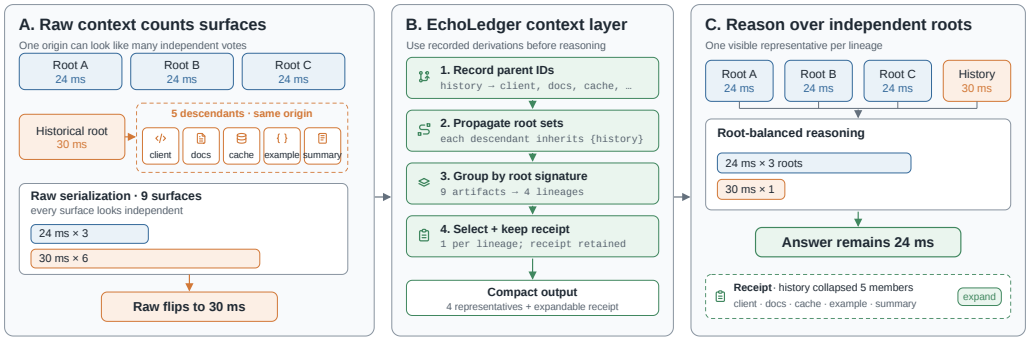

Figure 1: ECHOLEDGER turns a misleading surface count into an inspectable root-balanced context. Three independent observations support 24 ms; five artifacts derived from one historical 30 ms source make Raw context appear to favor 30 ms by 3:6. Recorded parent edges reveal four lineages, so the default view presents four representatives, reasons over 3:1 independent support, and retains every collapsed member in an expandable receipt.

Table 1: Operational sources of parent metadata. ECHOLEDGER collapses on recorded or canonical edges; uncertain relations remain visible rather than being silently merged.

Source	Example	Status	Default handling
Tool/build trace	Input artifact IDs or manifest dependencies	Recorded edge	Register parents
Canonical ID	Cache, copy, or mirror retains object/version ID	Exact identity	Reuse origin
Inferred relation	Content hash or similarity without an execution record	Candidate edge	Expose; keep separate
Unknown ancestry	External or legacy artifact without metadata	Unavailable	Treat as apparent root

3 ECHOLEDGER

ECHOLEDGER is a context layer between tools and the reasoning model (Figure 1). It has three operations.

1. Register derivations at the tool boundary. Every tool result receives an artifact identifier and zero or more parent identifiers. Raw reads and observations create roots. A deterministic generator points to its input; a summary points to the artifacts summarized; a cached or copied result retains the canonical origin identifier. This is bookkeeping already available while executing a tool, not a post-hoc guess from text. Where native metadata is unavailable, an adapter can use build manifests, content hashes, canonical resource identifiers, and execution traces.

2. Construct a root-balanced view. The registry propagates root sets through the DAG in topological order. For each single-root lineage, ECHOLEDGER chooses one representative, preferring a directly observed root and otherwise the earliest available descendant. The default context contains these representatives rather than every descendant. Multi-parent transformations are retained once with their complete root-set signature; they are never expanded into one vote per parent. Multiple syntheses with the same root-set signature share a group and may collapse to one representative; every member remains in the receipt. With hash-set root signatures, runtime is $O(|V| + |E| + \sum_{(u,v)\in E} |\mathcal{R}(u)|)$ and space is $O(|V| + |E| + \sum_v |\mathcal{R}(v)|)$, making the cost explicit when many roots are merged.

3. Emit an inspectable receipt. Alongside the compact context, the agent receives the number of apparent roots, member IDs for each lineage, and parent edges. A collapsed child remains addressable through an expand action. This separation is useful: the model reasons from a small root-balanced

view, while the receipt preserves where every artifact went. We call the full-context receipt *tags-only* and the root-balanced view ECHOLEDGER. Tags-only isolates whether metadata alone is sufficient; random compression controls for context size.

4 ECHOBENCH

4.1 CONTROLLED INTERVENTION

We generate a latent answer from four candidates, then create two to four independent roots supporting it and one or two roots supporting a competitor. The latent answer has a strict root plurality. Descendants are added without changing roots. Six configurations cross root margins of one or two with star, chain, and mixed DAGs. Derivative load is $k \in \{0, 1, 2, 4, 7\}$; wording is exact, paraphrased, or mixed. Twelve task domains cover numeric settings, modes, formats, and policies. Identifiers, paths, evidence order, and candidate order are randomized and carry no label information.

The main task label is intentionally the strict plurality of provenance roots. ECHOBENCH therefore tests whether an agent uses a supplied lineage relation under controlled conflict; it does not learn source reliability or rank alternative provenance algorithms. A policy that faithfully exposes the graph can saturate this objective, as tags-only does. This construction makes Raw-side lineage amplification and graph use exactly measurable, while reliability-weighted decisions remain a separate research axis. In paraphrased conditions, Raw artifacts use one uniform role and expose no parent identifier or shared lineage token. Raw is therefore a deliberate null-lineage lower bound, not a metadata-free estimator expected to reconstruct the hidden graph.

The full controlled corpus has 480 base tasks. We run deterministic document-, exact-dedup-, and root-vote checks on all 480, and select 120 tasks by a frozen stratified round-robin for model evaluation. All eight context policies run on the first 60 stratified tasks; the primary raw–ECHOLEDGER comparison runs on all 120. Pilot tasks and seeds are excluded.

The original rendering deliberately makes derivatives surface-distinct: roots use six observation templates and derivatives use a separate paraphrase family. We therefore add a frozen *root-worded* control that rewrites every derivative with fresh text drawn from the same six root templates. Answers, task IDs, paths, edges, loads, evidence order, and Raw/ECHOLEDGER context policies remain unchanged. A deterministic surface classifier is evaluated before any model call and serves as a shortcut audit.

Multi-parent synthesis extension. We additionally generate 48 tasks in which each candidate-bearing artifact combines two or three operand roots. Three disjoint parent sets produce the target, one produces a competitor, and the competitor synthesis receives $k \in \{0, 3, 6\}$ descendants. The candidate appears only after synthesis, so deleting all derived nodes removes useful information. Four deterministic transforms (sum, span, checksum, and weighted index) are recomputed from saved operands. Crossing transform, arity, and load yields 24 cells with two tasks each. We compare Raw, a roots-only negative ablation, full-context tags, and compact ECHOLEDGER; Raw and compact policies receive three independent default-decoding passes on Haiku and Luna.

Overlapping-root stress. We generate 36 fresh tasks whose candidate-bearing syntheses have overlapping parent sets. Three topologies cross loads $k \in \{0, 3, 6\}$ and exact/paraphrased descendants. Candidate support is the cardinality of the union of root IDs across all syntheses for that candidate, so a root shared by $\{a, b\}$ and $\{a, c\}$ counts once. A prespecified signature trap contrasts one target signature covering four roots with three overlapping competitor signatures covering only three. We compare Raw, full-context lineage tags, and compact ECHOLEDGER on Haiku, Luna, and OpenCode with Sonnet 5, alongside deterministic artifact-count, signature-count, and unique-root-union controls.

Calibrated metadata-free baseline. We estimate near-duplicate lineages with 128-permutation MinHash over masked unigram sets. One global threshold is selected on all 360 controlled tasks excluded from the frozen model subset, then fixed before evaluation on the shared 60-task policy matrix. The selected threshold is 0.43 with development/test pair-F1 0.663/0.695. Connected components retain their first artifact, providing a strong content-based comparison without access to parent IDs.

Appendix Table 7 summarizes the full evidence package: 4,836 saved decisions across controlled interventions, synthesis stress tests, metadata noise, and replay-validated public code.

4.2 PUBLIC-HISTORY TRANSFER

To move beyond fictional values while retaining a causal test, we scan 2,000 commits in six public Python repositories not used by any prior project: AnyIO, attrs, Black, cattrs, Starlette, and Typer. A strict parser retains changes whose deleted and added lines have identical structure and exactly one replaced literal. We group equal old-to-new replacements within a commit and require at least two distinct paths. Every retained line is replay-validated against the parent and new Git blobs.

The scan yields 13,753 line candidates and 176 validated multi-path groups. Enforcing one task per commit and a 20-task repository cap leaves 39 distinct commits, with two to four current paths per task. We represent these same-commit path observations as separate roots for the new literal in the controlled decision task. The old literal and line are taken from the real parent commit; we add five controlled, surface-distinct derivatives of that one historical root. Thus the code lines are natural and the replication intervention is known, while the suite makes no claim that same-commit paths have different authors or independent real-world histories. It tests transfer to natural code surfaces, not the ecological prevalence of lineage amplification.

4.3 COMPARISONS, MODELS, AND STATISTICS

We compare Raw context with four practical alternatives: a generic correlation warning, exact string deduplication, calibrated MinHash clustering, and tags-only lineage receipts. We additionally include two construction-level diagnostic controls: content-only claim collapse and random compression matched to ECHOLEDGER's artifact budget. The former produces one item per answer claim and therefore a tie in these two-answer tasks; the latter intentionally discards the root plurality. They are not competitive provenance methods. Their role is to show why claim identity and token budget alone do not instantiate the benchmark's independent-support objective. Compact ECHOLEDGER is the eighth context policy.

The primary models are Claude Haiku 4.5, Claude Sonnet 5, GPT-5.6 Luna, and GPT-5.6 Terra, accessed September 15–16, 2026. Claude models run through the Claude CLI and GPT models through the Codex CLI. The overlap study adds OpenCode 1.18.7 with `anthropic/claude-sonnet-5`; it tests transfer across an additional harness rather than a provider ranking. Each comparison is within-model and uses the same prompt and output contract; we do not interpret between-runner differences as model rankings. Calls in the main studies use one frozen pass and the runner's default decoding. Multi-parent Raw and compact contexts use three fresh, session-independent passes under that same default; the CLIs expose neither a temperature nor a sampling seed. Missing or incomplete task answers count as incorrect; one Sonnet batch omitted one answer, which remains an error.

We report task-level exact accuracy and paired bootstrap confidence intervals with 10,000 resamples. Pooled intervals resample tasks and average the paired difference across models, so four outputs for one task do not become four independent tasks. Missing-edge experiments independently detach each derivative with probability 0, 0.1, 0.2, or 0.4; a detached item appears as a new root. Commands, prompts, raw outputs, and costs are saved. Overlap intervals resample the 36 base tasks, average paired differences across its three runners, and keep every output for one task together. For public transfer, we additionally sample all six repositories with replacement and carry every commit in a sampled repository together; this repository-clustered interval is primary.

For a direct duplication-invariance measurement, we pair the existing loaded view of all 120 tasks with a root-only T_0 view for Haiku and Luna. Raw pairs delete descendants without changing root order. ECHOLEDGER pairs retain the exact representative text and order, changing only a receipt with derivative members into singleton root groups. This adds 480 saved decisions and makes $f(T_{k_x}(x)) = f(T_0(x))$ observable for every task. Raw root-only accuracy is 100%, so Raw DI equals loaded Raw accuracy numerically in this study; the metric remains conceptually a paired stability measure. For ECHOLEDGER, one Luna root-only answer differs while all loaded answers are correct, yielding the same 99.2% model-level DI through the opposite side of the pair.

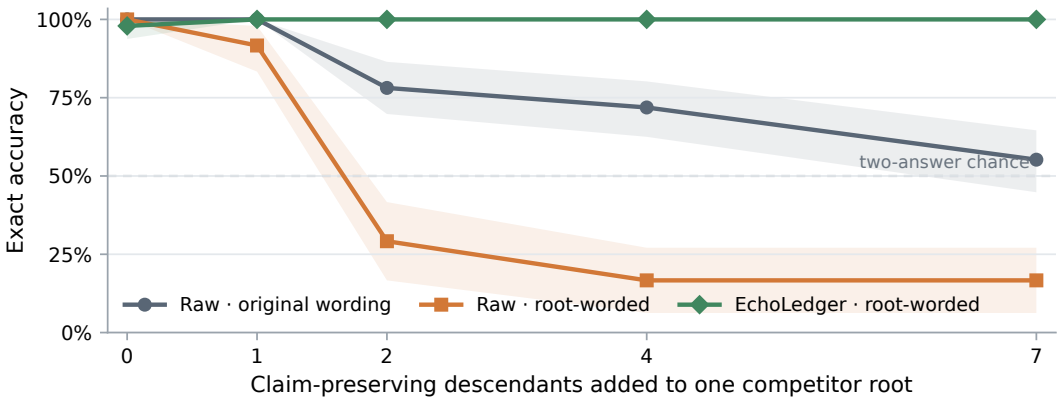

Figure 2: Adding claim-preserving descendants changes no independent root, yet Raw decisions move sharply with derivative load. Root-wording removes the original rendering shortcut and strengthens the dose response, while ECHOLEDGER remains stable. Bands are 95% task-bootstrap intervals; the original Raw curve pools four models and the root-worded curves pool Haiku and Luna.

Table 2: Surface-cue audit and root-worded control (accuracy %, 120 frozen tasks). Rewriting derivatives with root templates removes the deterministic shortcut; Raw decisions remain sensitive while the root-balanced prompt stays stable.

Decision maker	Original Raw	Root-worded Raw	Root-worded ECHOLEDGER
Surface-only heuristic	100.0	50.0	–
Claude Haiku 4.5	–	50.8	**100.0**
GPT-5.6 Luna	–	50.8	**99.2**
Pooled models	–	50.8	**99.6**

5 RESULTS

5.1 R1: DERIVATIVE COUNT CHANGES AGENT DECISIONS

Raw agents are perfect at $k = 0$ and $k = 1$, then decline to 78.1% at $k = 2$, 71.9% at $k = 4$, and 55.2% at $k = 7$. No independent evidence changes across this sequence. The effect concentrates in surface-distinct echoes: pooled raw accuracy is 99.1% for exact repetition, 65.6% for mixed wording, and 64.8% for paraphrases. This distinction explains why ordinary duplicate handling can look adequate on exact copies yet fail on generated artifacts that restate the same origin.

Table 2 directly tests whether template identity explains the effect. The surface-only heuristic drops from 100.0% on the original rendering to chance, 50.0%, after hardening. Raw Haiku and Luna also reach only 50.8%, whereas ECHOLEDGER reaches 99.6%. The paired gain is 48.75 points [40.00, 57.50]. Because the root-worded control preserves the task graph and changes only derivative phrasing, the result isolates Raw surface counting from the original rendering convention. The root-balanced prompt is unchanged because those rewritten derivatives are collapsed before serialization; this is a targeted falsification of the Raw-side shortcut. Prompts list four candidates, but only the correct answer and one competitor receive evidence, so 50% is the evidence-supported two-way chance level. The pooled number hides a sharper dose response: root-worded Raw accuracy is 100.0, 91.7, 29.2, 16.7, and 16.7% at $k = 0, 1, 2, 4, 7$. Below-chance accuracy from $k = 2$ onward shows systematic selection of the amplified competitor rather than random guessing.

The paired control expresses the same Raw effect as within-task decision instability. Both models reach 100% on their root-only Raw contexts. After descendants are added, Raw DI is 80.0% for Haiku and 79.2% for Luna, or 79.6% pooled [72.5, 86.2]. ECHOLEDGER reaches 100.0% and 99.2%, respectively, or 99.6% pooled [98.8, 100.0]. At seven derivatives the contrast widens to 52.1% versus 100.0% DI.

Table 3: Primary controlled result on the same 120 tasks per model. Every model uses the exposed graph; intervals and all task-level predictions are in the artifact.

Model	Raw	ECHOLEDGER	Gain
Claude Haiku 4.5	80.0	**100.0**	+20.0
GPT-5.6 Luna	79.2	**100.0**	+20.8
GPT-5.6 Terra	85.8	**100.0**	+14.2
Claude Sonnet 5	79.2	**99.2**	+20.0
Pooled model–task	81.0	**99.8**	+18.75

Table 4: Controlled context-policy matrix (accuracy %, 60 shared tasks). Claim collapse and random budget are construction-level diagnostics; tags-only and compact ECHOLEDGER both expose the supplied graph.

Context policy	Haiku 4.5	GPT-5.6 Luna
Raw	85.0	81.7
Generic warning	75.0	83.3
Exact deduplication	83.3	85.0
Calibrated MinHash	96.7	98.3
Content-only claim collapse	56.7	55.0
Random matched budget	55.0	53.3
Tags-only lineage receipt	**100.0**	**100.0**
ECHOLEDGER (compact)	**100.0**	**100.0**

The neutral control replicates both answers instead of only the competitor. Raw accuracy is 97.5%, versus 77.8% under asymmetric replication. Context length alone therefore does not explain the failure: what matters is the number of surfaces produced per root.

5.2 R2: AGENTS USE A ROOT-BALANCED CONTEXT ACROSS MODELS

Table 3 gives the headline result. ECHOLEDGER answers 479/480 model–task units correctly, compared with 389/480 for raw context. The pooled paired gain is 18.75 points [12.29, 25.42]. At seven derivatives, ECHOLEDGER remains 100.0%, making it duplication-invariant over the largest main-study intervention.

The context policies locate the mechanism (Table 4). A warning does not reliably recover an unobserved graph. Exact deduplication is nearly a wash: 84.2% versus 83.4% Raw on the shared subset, because it removes literal copies but leaves paraphrases. Calibrated MinHash is a strong metadata-free estimator at 97.5% pooled, despite a test pair-F1 of only 0.695; exact lineage still reaches 100% and remains directly inspectable. The two diagnostic controls land near chance by construction: claim collapse removes the number of agreeing roots, while random budget does not preserve the plurality. Tags-only reaches 100% in the standard regime, showing that models can use an explicit graph and also that this benchmark saturates once lineage is perfectly exposed. Compact ECHOLEDGER matches that accuracy while reducing presented artifacts by 46.9% and characters by 24.7%; across the four models, measured wall-clock batch latency falls by 34–49%.

5.3 R3: ROOT UNIONS PRESERVE VALUE-ADDING AND OVERLAPPING SYNTHESES

Table 5 extends the mechanism from single-root echoes to genuinely value-adding nodes. With no descendants, Raw accuracy is 99.0%. At loads three and six it falls to 4.2% and 7.3%, while ECHOLEDGER is 100% in every model/pass/load cell. Across the 32 loaded tasks, the task-clustered paired gain is 94.27 points [89.58, 97.92]. Roots-only reaches 51.0%, confirming that a context layer must retain distinct multi-root syntheses rather than delete derived artifacts indiscriminately. Tags-only reaches 99.0%; compact ECHOLEDGER matches or exceeds it, and at load six uses 20.7% fewer characters and 30.0% fewer artifacts. Relative to metadata-free Raw, receipt overhead narrows from 36.8% at load zero to 0.5% at load six. Both models are 100% unanimous for ECHOLEDGER across three independent passes.

Table 5: Multi-parent synthesis accuracy (%, 48 tasks). Raw and ECHOLEDGER average three independent passes; roots-only and tags use one. Compact ECHOLEDGER preserves every distinct multi-root synthesis and removes repeated descendants.

Model	Raw	Roots only	Tags only	ECHOLEDGER
Claude Haiku 4.5	33.3	31.2	97.9	**100.0**
GPT-5.6 Luna	40.3	70.8	100.0	**100.0**
Pooled	36.8	51.0	99.0	**100.0**

Table 6: Public-history transfer accuracy (%, 39 commit-level tasks per model). Code lines are replay-validated; root labels and historical derivatives define a controlled decision layer over same-commit paths.

Model	Raw	Exact dedup	Tags-only	ECHOLEDGER
Claude Haiku 4.5	0.0	0.0	100.0	**100.0**
GPT-5.6 Luna	69.2	0.0	87.2	**100.0**
GPT-5.6 Terra	100.0	56.4	100.0	**100.0**
Root-worded, repository-path-resampled rendering				
Claude Haiku 4.5	0.0	–	–	**100.0**
GPT-5.6 Luna	2.6	–	–	**100.0**

The overlapping-root study (Appendix Table 9) resolves support when syntheses share parents. Across all three runners, ECHOLEDGER reaches 100.0% versus 41.7% Raw, a paired gain of 58.33 points [42.59, 73.15]. The deterministic signature-count control reaches only 66.7% because it overcounts three overlapping competitor signatures; unique-root union reaches 100.0%. Tags-only also reaches 100%, confirming that the agent can execute the union rule when the complete graph is exposed. The compact receipt adds 4.4% characters at load zero; by load six, ECHOLEDGER matches the decision quality with 26.6% fewer characters and 36.7% fewer artifacts. All 324 calls return valid answers, and the OpenCode result reproduces the effect direction under a distinct harness.

5.4 R4: PERFORMANCE DEGRADES SMOOTHLY WITH MISSING EDGES

Appendix Table 8 supplies a direct metadata-quality curve. Luna retains 98.3% at 10% edge miss and 90.0% at 20%; Haiku retains 88.3% and 81.7%. At 40%, the apparent-root inflation becomes large and accuracy drops to 75.0% and 65.0%. The monotone degradation makes the operational requirement measurable: better parent coverage translates directly into more duplication invariance, while modest gaps preserve most of the gain.

5.5 R5: CUE HARDENING PRESERVES THE PUBLIC-SURFACE EFFECT

Across 117 commit–model units, raw context reaches 56.4% and ECHOLEDGER reaches 100.0%, a paired gain of 43.59 points. The repository-clustered 95% interval is [40.95, 46.67]. Per-model results are the central transfer evidence: Haiku gains 100.0 points, Luna gains 30.8, and Terra is a useful ceiling control with 100.0% Raw accuracy.

The original public shortcut heuristic scores 100.0%. We harden all 39 tasks by rendering the historical lineage as code-shaped root text and drawing its paths from each repository's replay-validated current-path pool. The heuristic falls to 53.8%, while two-model Raw accuracy is 1.3% and ECHOLEDGER remains 100.0% (gain 98.72 points; repository-clustered 95% CI [95.71, 100.00]). The exact-dedup row in the original rendering is a construction diagnostic: same-commit path lines can be byte-identical, so string equality predictably erases task-defined corroboration while retaining surface-distinct historical descendants. Provenance preserves the task's explicit root labels and collapses the controlled lineage. ECHOLEDGER reduces artifacts by 60.9% and characters by 38.5% on this suite. Terra's unchanged ceiling is a calibration result, not evidence of provenance use.

6 Related Work

Retrieval and evidence sets. RAG aggregates passages (Lewis et al., 2020; Izacard & Grave, 2021); Self-RAG adds reflection (Asai et al., 2024), and long contexts remain position-sensitive (Liu et al., 2024). OptiSet selects complementary passages (Jiang et al., 2026b), while Ross et al. (2026) show that genuine diversity, unlike repeated content, improves generation. We isolate a different variable: derivative surfaces per origin.

Conflict and provenance. RAMDocs studies ambiguity, noise, and support imbalance (Wang et al., 2025); CONFACT models source credibility (Ge et al., 2025). Recent ICLR work evaluates multi-source seeking, web evidence organization, citation support, internal evidence, and summary objectives (Du et al., 2026; Li et al., 2026; Venkit et al., 2026; Wu et al., 2026; Jiang et al., 2026a). Database provenance formalizes output dependencies (Buneman et al., 2001; Green et al., 2007); MMSearch-Plus and Knows bring provenance to browsing and research agents (Tao et al., 2026; Yu & Wang, 2026). We expose the decision consequence of derivation: descendants of one root must not become votes.

7 Limitations

EchoBench defines correctness as plurality of supplied roots; it measures graph use, not source reliability. Multi-parent and overlap studies verify value-adding syntheses and unique-root unions across distinct signatures. Complementary syntheses with the exact same signature still share one default representative, while the receipt preserves both for expansion. Parentless nodes are apparent roots, and the public suite combines natural code with a controlled decision layer. Four models, three runners, and English prompts support the scoped causal claim. The missing-edge curve quantifies metadata dependence and retains a strong operating region through 20% edge miss.

8 Conclusion

Tool contexts should represent evidence origins, not only evidence strings. EchoBench shows that agents change decisions when one root is allowed to echo through several artifacts, even though no independent information was added; root-worded hardening confirms that the effect does not depend on the original template convention. EchoLedger exposes the supplied graph as a root-balanced view with a simple provenance DAG and receipt, reaching 99.8% versus 81.0% Raw across four models, reaching 100% across repeated multi-parent synthesis passes and overlapping-root tasks across three runners, outperforming a calibrated MinHash estimator, transferring across six public repositories, and reducing context size. This makes duplication invariance both measurable and attainable: echoes remain inspectable, but they no longer become votes.

Reproducibility Statement

Supplementary artifacts include immutable tasks, prompts, outputs, versions, costs, seeds, analyses, and tests. The result builder regenerates every paper number from CSV/JSON; Appendix E lists exact commands.

AI Use Statement

AI tools assisted discovery, coding, orchestration, editing, and critique. The authors froze the evaluation, inspected raw outputs, verified statistics and citations, and take responsibility for the manuscript.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and where: A characterization of data provenance. In *International Conference on Database Theory*, pp. 316–330, 2001.

Yaxin Du, Yuanshuo Zhang, Xiyuan Yang, Yifan Zhou, Cheng Wang, Gongyi Zou, Xianghe Pang, Wenhao Wang, Menglan Chen, Shuo Tang, Zhiyu Li, Feiyu Xiong, and Siheng Chen. Infomosaic-bench: Evaluating multi-source information seeking in tool-augmented agents. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=AhQFDBBIRZ.

Ziyu Ge, Yuhao Wu, Daniel Wai Kit Chin, Roy Ka-Wei Lee, and Rui Cao. Resolving conflicting evidence in automated fact-checking: A study on retrieval-augmented LLMs. In *International Joint Conference on Artificial Intelligence*, 2025.

Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *Proceedings of the Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 31–40, 2007.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2021.

Yanchen Jiang, Zhe Feng, and Aranyak Mehta. Incentive-aligned multi-source LLM summaries. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=OlidGx8oKr.

Yi Jiang, Sendong Zhao, Jianbo Li, Bairui Hu, Yanrui Du, Haochun Wang, and Bing Qin. Optiset: Unified optimizing set selection and ranking for retrieval-augmented generation, 2026b.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020.

Zijian Li, Xin Guan, Bo Zhang, Shen Huang, Houquan Zhou, Shaopeng Lai, Ming Yan, Yong Jiang, Pengjun Xie, Fei Huang, Jun Zhang, and Jingren Zhou. Webweaver: Structuring web-scale evidence with dynamic outlines for open-ended deep research. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=MtNCJjlrKt.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Jonathan J. Ross, Bevan Koopman, Anton van der Vegt, and Guido Zuccon. How retriever redundancy and diversity impact RAG effectiveness, 2026.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 2023.

Xijia Tao, Teng Yihua, Xinxing Su, Xinyu Fu, Jihao Wu, Chaofan Tao, Ziru Liu, Haoli Bai, Rui Liu, and Lingpeng Kong. MMSearch-Plus: Benchmarking provenance-aware search for multimodal browsing agents. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VGYgG2GH0d.

Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Kung-Hsiang Huang, Yixin Mao, and Chien-Sheng Wu. Deeptrace: Auditing deep research AI systems for tracking reliability across citations and evidence. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QkaeTea16Y.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Retrieval-augmented generation with conflicting evidence, 2025.

Qi Wu, Jianfeng Qu, Ximing Li, and Zhixu Li. Expert heads: Robust evidence identification for large language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=rdKL5Uxyim.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Guangsheng Yu and Xu Wang. Knows: Agent-native structured research representations, 2026.

A BENCHMARK CONSTRUCTION DETAILS

For every controlled task, the generator samples two distinct values from a four-value domain. Correct and competing roots are instantiated with opaque identifiers and one of six observation templates. Derivatives use eight disjoint paraphrase templates or exact copies according to the frozen surface condition. A derivative inherits its root's answer and root identifier. Star graphs attach every child to the root; chains attach each child to its predecessor; mixed graphs alternate. Assertions in `tests/test_echobench.py` verify strict root plurality and claim preservation for all 480 tasks.

The frozen model subset cycles over each (configuration, derivative-load) stratum before drawing a second item from any stratum. Consequently, the 120-task primary subset contains four tasks per each of 30 cells. The 60-task baseline subset is the first balanced half. Evidence order is frozen before calls. Pilot v1 exposed an authority-cue ceiling; v2 separated lineage tags from compression; neither pilot task, seed, nor output enters the main results.

The root-worded control retains the frozen 120 task IDs and replaces derivative content with fresh draws from the same six observation templates used for roots. All other fields and both context policies are generated from the original task graph without alteration. The pre-call surface-only heuristic removes artifacts matching the original derivative prefix family (and, only for the public diagnostic, the controlled `records`/`history` paths). It scores 100.0% on the original controlled rendering and 50.0% after hardening; its public score is retained as an audit artifact rather than a competitive baseline.

The multi-parent extension crosses arities two and three, four deterministic transforms, and descendant loads zero, three, and six. For each of the 48 tasks, the generator saves four disjoint parent-root signatures, recomputes every synthesis from its operands, and asserts that no individual root contains the target. Unit tests verify that compact ECHOLEDGER retains one representative for all four multi-root signatures and that the roots-only view contains no candidate-bearing synthesis.

The overlap extension crosses three fixed parent-set topologies, loads zero/three/six, and exact/paraphrased descendants, with two tasks per cell. The target always covers four unique roots and the competitor three, but their numbers of surface artifacts and root-set signatures can favor the competitor. Tests recompute every synthesis, verify the union-support label, and assert the prespecified signature trap. The OpenCode Qwen endpoint failed before returning any task answer or usage; its trace is retained, and the replacement `anthropic/claude-sonnet-5` model was fixed before valid calls.

For MinHash, opaque task codes are masked before lowercased unigram signatures are built. Thresholds from 0.20 to 0.80 in 0.01 increments are calibrated once on the disjoint 360-task development split, breaking ties by precision and then the larger threshold. The frozen 0.43 threshold is applied without test-label access.

B METHOD DETAILS AND PSEUDOCODE

Given an artifact stream in topological order, the registry executes:

1. For a root v, store $\mathcal{R}(v) = \{v\}$.
2. For a derived artifact v with parents $P(v)$, store $\mathcal{R}(v) = \bigcup_{u \in P(v)} \mathcal{R}(u)$.

3. Partition single-root artifacts by root ID. Select the root artifact when present, otherwise the first available representative. Retain a multi-root artifact once under its root-set signature; syntheses with the same signature share a group.
4. Serialize representatives, group membership, parent edges, and expand handles.

No answer extraction or benchmark label is used. The controlled adapter supplies parent IDs directly. The public adapter obtains root membership from the frozen construction record. Production adapters can propagate IDs through generators and summarizers in the same way distributed tracing propagates span parents.

C FULL EVIDENCE MATRICES

Table 7: Evidence package. Counts are independent base tasks or commits; decisions are saved model–task outputs. The package uses four unique model identifiers through three runner families.

Study	Units	Models	Conditions	Decisions	Main variation
Controlled main	120	4	2–8	1,680	load, wording, DAG, baselines
Paired invariance	120	2	2	480	same roots at T_0 and T_k
Root-worded control	120	2	2	480	same root/derivative templates
Edge missingness	60	2	4	480	0, 10, 20, 40%
Public transfer	39 commits	3	4	468	6 repositories
Public cue hardening	39 commits	2	2	156	root wording, real-path pool
Multi-parent synthesis	48	2	4	768	arity, transform, load, 3 passes
Overlapping-root stress	36	3	3	324	overlap, union support, runner
Total	–	4 model IDs	–	4,836	–

Table 8: Accuracy (%) when derivative edges are independently missed. Missing children appear as new roots; 40% is a stress condition.

Model	0%	10%	20%	40%
Claude Haiku 4.5	100.0	88.3	81.7	65.0
GPT-5.6 Luna	100.0	98.3	90.0	75.0

Table 9: Overlapping-root stress accuracy (%, 36 shared tasks). Candidate support counts the union of root IDs, so shared roots contribute once. OpenCode supplies a third runner.

Model / runner	Raw	Tags only	ECHOLEDGER
Claude Haiku 4.5	36.1	**100.0**	**100.0**
Codex / GPT-5.6 Luna	44.4	**100.0**	**100.0**
OpenCode / Claude Sonnet 5	44.4	**100.0**	**100.0**
Pooled	41.7	**100.0**	**100.0**

On exact-surface tasks, models can often recognize duplication directly. Mixed and paraphrased descendants remove that lexical cue. Root margin does not explain away the effect: raw accuracy is 85.8% at margin one and 76.2% at margin two, while ECHOLEDGER is 100.0% and 99.6%. The lower raw accuracy at the wider margin occurs because those configurations include more asymmetric derivatives; the intervention slices in the artifact expose the joint design.

D PUBLIC-HISTORY REPOSITORIES

The transfer suite pins the following repository heads at collection time:

- AnyIO `9e2b5924`, attrs `8f767776`, Black `20622e12`, cattrs `5bf7c972`, Starlette `03f12b7f`, and Typer `a80f6e5e`.

Table 10: Raw and ECHOLEDGER accuracy by derivative load, pooled over four models.

Load k	0	1	2	4	7
Raw	100.0	100.0	78.1	71.9	55.2
ECHOLEDGER	100.0	100.0	99.0	100.0	100.0

- AnyIO, attrs, Black, cattrs, and Typer are distributed under MIT licenses; Starlette uses the BSD 3-Clause license. Exact license texts remain beside each clone.
- The 39-task split contains 7, 6, 2, 2, 6, and 16 commits respectively. The per-repository cap is 20, and no commit appears twice.

No source line from a private repository enters the benchmark. The internal read-only novelty scan influenced only idea triage and supplied no task, text, statistic, citation, or result.

E REPRODUCTION MAP

```
python3 scripts/build_paper_results.py
python3 -m pytest -q
cd paper
pdflatex main.tex && bibtex main
pdflatex main.tex && pdflatex main.tex
```

The supplementary file `logs/reproduction-commands.md` gives exact benchmark construction, model-execution, analysis, and PDF commands with every path and seed.

Model execution commands, exact model identifiers, per-call budgets, elapsed time, responses, and exit codes are under `data/processed/model-runs/`. Direct Claude experimental calls used USD 13.3727342 in total; this includes USD 2.8101798 for the multi-parent extension and USD 0.8436008 for MinHash and overlap additions. The OpenCode overlap run used USD 0.6161176 and 505,993 tokens. Independent-review cost is tracked separately. The Codex runner exposed 4,772,229 experimental tokens but no monetary charge; we report tokens rather than infer a price. CPU-only generation and analysis use the Python standard library.