LINEAGE LEDGERS: CREDITING SOURCES ONCE, AND ONLY WHERE THEY LOOKED

Anonymous authors
Paper under double-blind review

ABSTRACT

Counting passages is a cheap way to aggregate evidence, and it hands the outcome to whoever copies most, since copying a passage costs nothing while observing the world does not. Confidence here is credited to lineage-distinct sources rather than to deliveries, and each source is credited only along the directions its observation constrains. Truth discovery and pedigree methods instead treat a source as a single vote. A ledger realizes both parts, through an additive union across sources, a max-register join within a source, and a per-source precision matrix. Its credit comes out invariant to redelivery and to arrival order. A coordinated bottom-k sketch bounds the storage, with a finite-sample error bound that holds for scalar and for directional credit alike and does not grow with the number of deliveries. Two pre-registered experiments test the principle. A frozen 7B-parameter reader loses 0.1167 gold-answer recall when one false paragraph is copied sixteen times. Reading from the ledger recovers the whole loss on one fifth of the context, whereas MinHash, the deployed defence, deletes the true paragraph in half of all records when the falsehood arrives first. On real MuSiQue retrieval, support kept per requirement predicts whether retrieved evidence contains every needed fact 0.1535 AUROC better than the same support averaged into a scalar, which sits at chance within every hop count. Inside a learned multi-hop memory the ledger holds its clean accuracy under sixteenfold redelivery, exact or re-chunked, while the undefended memory loses more than a quarter of it, as does every near-duplicate detector once the copies are re-chunked. When only fragments of a source arrive, the ledger matches reading every fragment while crediting one source instead of sixteen.

1 INTRODUCTION

Language models tend to side with whichever answer more of their context supports. When retrieved passages disagree, Xie et al. (2024) find that models follow the majority, and the passages in their study are distinct texts. A majority on the open web is often manufactured rather than observed. One report is syndicated across outlets, one paragraph is re-chunked by a retriever, and one message is retransmitted by a network. Copying is free while observation is not, so a reader that counts passages rewards whoever copies most.

This work starts from the principle that repetition is not corroboration. Confidence in a claim should grow with the number of lineage-distinct sources that support it, and not with the number of times those sources are delivered. The principle does not ask a memory to discard repeated material, since a later version of a source can be a better representation of it. Two quantities that message-passing systems routinely conflate are therefore kept apart. *Computational innovation* measures how much a message improves the current representation, and it is left unconstrained. *Evidence credit* measures how much source-level support a message contributes, and lineage bounds it. Ten transformations of one observation are not ten observations.

Truth discovery and pedigree methods already discount copies (Section 2). Beyond them, this work contributes the following.

- **A lineage ledger with guarantees.** Payloads add across lineage-distinct sources, and within a source a later version replaces an earlier one under a key whose mass term stops a

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.

redelivery from raising confidence. Credit is invariant to redelivery and to arrival order, and a bottom-k sketch bounds storage with an error bound that does not grow with deliveries (Section 4).

- **Directional credit.** A source earns credit only along the directions its observation constrains. A scalar source count, as used by truth discovery and pedigree methods, is its trace surrogate and claims support along directions no source observed, and the guarantees carry over (Proposition 1).
- **Two pre-registered demonstrations.** Copies of one false source move a frozen 7B reader monotonically, extending the majority-following of Xie et al. (2024) and Naphade (2026) to majorities manufactured by copying. The ledger removes the effect on one fifth of the context, while near-duplicate detection deletes the truth when the falsehood arrives first. On real retrieval, support kept per requirement predicts whether evidence suffices where the same support averaged into a scalar does not.
- **A learned memory.** The ledger holds clean accuracy under exact and re-chunked redelivery, where no defence and every near-duplicate detector lose more than a quarter of it. Given fragments alone it matches reading every fragment at one sixteenth of the credit. It costs less than no defence and conserves credit unchanged on three further backbones.

The term *lineage-distinct* stands in for *independent* throughout, since source identity alone cannot establish statistical independence.

2 Related Work

Work bearing on repeated evidence falls into four groups, which differ in what they treat as the unit of redundancy. The experimental comparisons in Section 6 draw their baselines from all four.

2.1 Duplicate detection

The standard defences operate on payloads rather than on lineage. MinHash over shingles (Broder, 1997) estimates resemblance from hashed n-grams. SimHash (Charikar, 2002) compares at a Hamming radius, and web-scale crawling fixes that radius by convention (Manku et al., 2007). Corpus cleaning applies exact and approximate matching at scale (Lee et al., 2022), and MinHash remains the deduplication stage of current web-scale pretraining pipelines (Penedo et al., 2024), so the comparison in Section 6 runs against deployed practice. Every method in this group asks whether two texts look alike, whereas lineage asks whether two texts descend from one observation. The two questions diverge in both directions. Overlapping chunks of one paragraph are not near-duplicates of one another, and a false paragraph that edits one span of a true one is a near-duplicate of it. Fabricating identity in order to defeat a lineage ledger is the Sybil attack in this setting (Douceur, 2002).

2.2 Evidence fusion

A second group fuses estimates rather than filtering them. Dempster's rule of combination (Shafer, 1976) is commutative and associative but not idempotent, so combining one source with itself sharpens the resulting belief. Covariance intersection (Julier & Uhlmann, 1997) fuses under convex weights and therefore stays consistent for any unknown cross-correlation, but it bounds confidence without recording which sources contributed, so it cannot credit genuine corroboration.

2.3 Source dependence and pedigree

The principle that copies are not corroboration is established in two other literatures. Truth discovery in data integration discounts values that one source copied from another, and detects copying from the errors that sources share (Dong et al., 2009a;b; 2010). Distributed fusion calls the same failure data incest and avoids it by propagating pedigree, the tags of the raw measurements behind each estimate (Arambel, 2008), most recently as products of primes whose least common multiple merges source sets without double counting (Chang et al., 2025). Both treat a source as a vote or as a set member. The ledger differs in three respects. It credits a source only along the directions its

observation constrains. It keeps for each source a versioned payload that a later arrival may refine without adding credit, established formally here (Corollary 1) rather than measured. It bounds its state with a mergeable sketch, where exact pedigree grows with the number of sources.

2.4 Message passing and belief propagation

A third group addresses cyclic feedback. Loopy propagation becomes overconfident on cycles (Murphy et al., 1999), and non-backtracking variants remove short cycles (Park et al., 2024). Duplication is a distinct failure, since a duplicated root inflates confidence on a tree where no loop exists. Idempotent merging of replicated state is standard in distributed systems (Shapiro et al., 2011), and duplicate-insensitive summaries combine sketches with multipath communication (Nath et al., 2004). The question added here concerns versioned learned payloads and the error a compressed payload incurs.

2.5 Retrieval and knowledge conflicts

A fourth group aggregates retrieved passages without source accounting (Lewis et al., 2020), and models favour the answer with more support (Xie et al., 2024). A handful of crafted passages can control a reader's answer (Zou et al., 2025), and the closest defence certifies quality against a bounded number of injected passages (Xiang et al., 2026), a budget that copies of one source exhaust for free. Paraphrasing one argument can persuade a model more than distinct support (Naphade, 2026), the grouping of retrieved records controls the weight a model gives them (Liu et al., 2026), and graph agents treat repeated paths to one observation as corroboration (Wang et al., 2026). These studies measure the failure, and the ledger is a mechanism that removes it.

3 Method

3.1 An arrival, and where its lineage comes from

Every message entering the memory is a triple. It carries a lineage identifier r, a payload, and a version. The identifier names the *observation* that the message descends from, not the message itself, so two arrivals share an identifier exactly when they report the same underlying observation, however far apart their surface forms.

The identifier is supplied with the arrival rather than inferred from its content. This is the central design decision and it is deliberate. Proposition 3 in Appendix A shows that no function of content can recover lineage, because m copies of one observation and m separate observations can induce the same distribution over text. The ledger therefore does not try to discover lineage. It conserves the lineage it is given, which is a different and achievable job. In practice the errors that sources share often reveal copying (Dong et al., 2009a), so a copy detector can supply the identifiers where provenance metadata is missing.

In the experiments the identifier is the identity of the supporting paragraph, so a verbatim copy, an overlapping re-chunking and a paraphrase of one paragraph all carry the same identifier, while a paragraph that contradicts it carries its own. A deployed retrieval system would instead use provenance metadata already present in the index, such as a document identifier or a canonical URL. What happens when an adversary fabricates identifiers is measured in Section 7 and Appendix G.

3.2 The ledger and the rule applied to each arrival

A cell holds a learned workspace and, separately, a ledger. The ledger is a map from each identifier it has seen to the best version of that observation it holds. Versions of one identifier are ordered by

$$\kappa(v) \;=\; \big(\rho_v,\; -m_v,\; h(z_v)\big), \tag{1}$$

where ρ_v is the refinement level, m_v the mass the version claims and $h(z_v)$ a hash of the payload. The middle term is what stops inflation. Two versions at equal refinement are equally informed about one observation, so the ledger keeps the one claiming *less* information, and a redelivery can move the belief without ever raising confidence. The hash makes the order total. In the text experiments an arrival claims no mass, so the key reduces to the version and the hash, and copies share version

zero, so the ledger keeps one rendering of each identifier, chosen by the hash. A window cut from a paragraph omits part of it and ranks below it.

Each arrival is then resolved by one rule, before it reaches the content path. If its identifier is absent from the ledger, it is admitted and recorded. If its identifier is present and its key strictly beats the incumbent, it is admitted and replaces the incumbent, which is refinement. Otherwise it is dropped and never reaches the model. Across identifiers the payloads add, and within one identifier a later version replaces an earlier one. Both operations are commutative, associative and idempotent, so the ledger reaches the same state under any arrival order, any delay and any number of redeliveries.

Credited evidence is then read off the ledger directly. It is one unit for each identifier the ledger holds, whatever the number of arrivals that carried it. That single sentence is the conservation property. The provenance-free control differs in exactly one respect, adding a unit per arrival rather than per identifier, which is why its credit tracks the delivery count.

3.3 Directional credit

Credit is not a count. Each source is credited with its own precision matrix,

$$\Lambda_i \;=\; \Lambda_{\text{prior}} + \sum_r \rho_{ir}\,\Lambda_r, \qquad \rho_{ir} \in [0,1], \tag{2}$$

so support accumulates only along the directions each observation constrains. The support a memory holds along a direction u is $u^\top \Lambda_i u$. Two lineage-distinct sources that both measure x corroborate x and say nothing about y, and a memory whose sources all constrain one direction stays at its prior along every other. A scalar count of sources, the form used by truth discovery, by grouping on document identifiers and by pedigree sets, amounts to replacing each Λ_r with $\frac{1}{d}\mathrm{tr}(\Lambda_r)\, I$. That keeps the total information and discards its direction, so it claims support along directions no source observed.

For text, a retrieved passage with embedding e_r contributes rank-one credit $e_r e_r^\top$. A question that needs several facts defines several requirement directions d_h, and the support along each is $s_h = \sum_r (d_h^\top e_r)^2$, summed over lineage-distinct passages. Evidence suffices only if every requirement is supported, so the directional confidence is the weakest of them, $\min_h s_h$, whereas a scalar confidence averages them and can be high while one requirement has no support at all.

How sources combine within one requirement is a separate choice. Summation measures corroboration and is exact when lineage-distinct sources also err independently. When they may share errors, or when the question is only whether some source covers each requirement, the conservative choice keeps the best single source, $\max_r (d_h^\top e_r)^2$, which is what covariance intersection (Julier & Uhlmann, 1997) yields along one direction. Both choices are directional, and only summation can register a second source. Section 6.3 measures both.

3.4 Bounded storage and what a reader sees

A bottom-k sketch (Cohen & Kaplan, 2007) bounds the ledger by retaining the k identifiers with the smallest coordinated hashes under a shared threshold θ, so two memories that never communicated retain the same subset and merge without double counting. Below k identifiers the sketch is exact, and above k the Horvitz-Thompson estimator (Horvitz & Thompson, 1952) $\hat{S} = \theta^{-1} \sum_{r \in S} \eta_r$ is unbiased (Appendix A). A reader, learned readout or frozen language model alike, reads the ledger state rather than the arrival stream, so it sees every lineage-distinct source once, in the best form that has arrived.

4 Theory

Compression of the interface means its approximation error needs to be bounded rather than assumed. Appendix A gives the assumptions and the proofs.

Theorem 1 (Categorical memory bound, informal)**.** *Fix a set of source payloads independent of the sketch hashes, and let the decoder satisfy Assumption 1. The KL divergence between the sketch-adapter model and the exact-ledger model then admits a finite-sample bound in terms of the repre-*

sentation error and the evidence-mass error. The bound depends on neither the number of deliveries nor the multiplicity of any root.

Corollary 1 (Invariance to redelivery)**.** *Let a delivery sequence contain N arrivals descended from R lineage-distinct roots, with $N \geq R$. Credited evidence, and therefore the reported confidence ceiling, depends on R and not on N. Redelivering a version already held leaves both unchanged however large N becomes, while a version that strictly beats the incumbent may still change the representation.*

An adversary who multiplies deliveries of a held source therefore gains nothing, and repeated deliveries change neither error term of Theorem 1. Proposition 2 prices the alternative for Gaussians sharing a mean, where crediting all N copies of one observation in d dimensions costs $\frac{d}{2}(N - 1 - \ln N)$ nats. The same guarantees hold for the directional credit of Equation 2.

Proposition 1 (Directional credit, informal)**.** *Let each held source carry a positive semidefinite precision as in Equation 2. (i) Support along every direction, and the weakest requirement's support, are functions of the ledger state and so inherit Corollary 1. (ii) Along any direction that no held source constrains, support equals the prior, whereas the scalar surrogate $\frac{1}{d}\mathrm{tr}(\Lambda_r)I$ adds $\frac{1}{d}\sum_r \rho_{ir}\mathrm{tr}(\Lambda_r)$ there while preserving the trace. (iii) Under the bottom-k sketch, with probability at least $1 - \delta$ the error in support is at most one radius uniformly over all directions, and hence in the weakest requirement's support. That radius depends on neither the number of deliveries nor the multiplicity of any root.*

For rank-one text credit over n held sources the radius is at most $n\sqrt{(n-k)/(k(n-1)\delta)}$, and the sketch never over-reports best-single-source support (Appendix A).

5 Experimental Setup

Two datasets, two frozen readers and one learned memory support the experiments below, and the subsections that follow fix the data, the three duplication settings, the baselines and the protocol.

5.1 Data and readers

MuSiQue (Trivedi et al., 2022) composes single-hop questions into chains of two to four hops, with explicit decompositions and per-step supporting paragraphs. AVeriTeC (Schlichtkrull et al., 2023; 2024) supplies a real-world verification setting with an official knowledge store. Passages for the learned models are encoded with a frozen BGE encoder (Xiao et al., 2024). The language-model experiments use Qwen2.5-7B-Instruct and Qwen2.5-1.5B-Instruct (Qwen et al., 2024) as frozen readers with greedy decoding. No reader is ever fine-tuned, and each arm differs from the others only in which passages its accounting rule admits to the context.

5.2 Syndicated misinformation

For each of the first 300 eligible two-hop MuSiQue records, a counterfactual paragraph is built by replacing the gold answer and its aliases in the final supporting paragraph with a substitute of the same kind, taken deterministically from a later record. The counterfactual carries its own lineage identifier and is delivered $m \in \{1, 2, 4, 8, 16\}$ times, as verbatim copies or as reworded copies framed by a distinct outlet label, arriving either after the true paragraphs or before them. A reply counts as a gold hit if it contains a gold answer or alias, and as misled if it contains the substitute. The predictions, endpoints and decision rules were committed before any reader output existed, and Appendix B reproduces them.

5.3 Directional credit

Two pre-registered experiments test Section 3.3, a synthetic one (Appendix C) and one on every answerable MuSiQue development record. The five of its twenty paragraphs most similar to the question form the retrieved set, which is sufficient when it holds every supporting paragraph. Requirement directions are the question's decomposition steps with every reference to an earlier answer removed, which stands in for a query decomposer and gives no score access to answers or support labels.

Table 1: What the results establish. Intervals are paired and cover 95%. Reader results are pre-registered and resampled over 300 records, learned memory results over five seeds.

Claim	Evidence	Section
Copies of one false source move a frozen reader	-0.1167 gold recall at sixteen copies, $[-0.1600, -0.0733]$	6.1
Lineage accounting removes the whole effect	$+0.1167$ recall, $[+0.0733, +0.1600]$, on one fifth of the context	6.1
Surface similarity cannot substitute for lineage	MinHash deletes the true passage in 52% of records	6.2
Corroboration is directional	$+0.1535$ within-hop AUROC over scalar support, $[+0.1312, +0.1755]$	6.3
Lineage helps confidence under copies	$+0.0481$ AUROC over provenance-free counting, $[+0.0294, +0.0673]$	6.3
Conservation protects a learned memory	$+0.0318$ top-1 over no defence at $\times 16$, $[+0.0168, +0.0468]$	6.4
Lineage protects where surface matching cannot	$+0.0307$ top-1 over the strongest detector, re-chunked $\times 16$, $[+0.0226, +0.0388]$	6.4
With fragments only, lineage bounds credit	credit 1.000 against 16.000, top-1 $+0.0064$, $[-0.0059, +0.0187]$	6.4
Conservation costs less than no defence	0.037 seconds against 0.063	6.5

5.4 Duplication regimes for the learned memory

Every hop of a chain is delivered at multiplicities of one to sixteen, as **exact** byte-identical copies, **re-chunked**, where the paragraph arrives once in full and then as distinct overlapping windows cut from it, or as **fragments**, where only the windows arrive. Every regime preserves the lineage identifier, making each a test of conservation rather than of identification. Appendix D constructs all three.

5.5 Baselines

Surface deduplication comprises canonical hashing as used in corpus cleaning (Lee et al., 2022), MinHash (Broder, 1997), SimHash (Charikar, 2002) at the published Hamming radius (Manku et al., 2007), and an embedding detector over BGE representations (Xiao et al., 2024) at a frozen cosine threshold. Evidential fusion comprises Dempster's rule (Shafer, 1976) and covariance intersection (Julier & Uhlmann, 1997). Non-backtracking message passing (Park et al., 2024) represents cycle suppression, and the provenance-free control processes every arrival. All detector thresholds are frozen literature defaults, and none is tuned on any evaluation population.

5.6 Protocol

Learned models run over a common set of five seeds, and differences are reported as exact paired t intervals over seeds. The learned memory is evaluated on $1{,}252$ two-hop development records ranked against 555 candidate answers. The frozen readers are deterministic, so their intervals come from $10{,}000$ paired bootstrap resamples over records. Credited evidence is reported relative to the lineage-distinct ceiling, which makes exact conservation read as 1.000 at every multiplicity.

6 Results

The results start at the reader and work outward, and Table 1 states what each one establishes.

6.1 Copies of one false source move a frozen reader

The most direct test asks whether a modern reader treats repetition as corroboration. Figure 1 and Table 2 report the pre-registered experiment, in which one counterfactual paragraph is delivered to Qwen2.5-7B-Instruct (Qwen et al., 2024) up to sixteen times. Appendix B reports every condition.

Without a defence the reader follows the copies. Its gold-answer recall falls from 0.5133 with one copy to 0.3967 with sixteen, a paired change of -0.1167, interval $[-0.1600, -0.0733]$, while the rate at which it reports the planted answer rises from 0.0100 to 0.1367. That decline is monotone in the number of copies, which extends the majority-following of Xie et al. (2024) to manufactured majorities.

Reading from the ledger removes the effect. All copies share one identifier, so the reader sees the counterfactual once at every multiplicity. At sixteen copies the ledger improves gold recall by $+0.1167$, interval $[+0.0733, +0.1600]$, which recovers the entire loss. Its context stays at 464.5 tokens against 2362.1, a 5.1-fold reduction. All five pre-registered predictions hold.

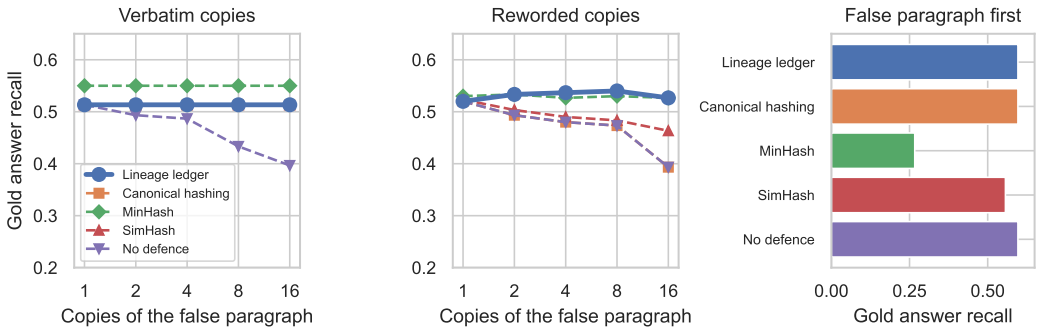

Figure 1: Gold-answer recall of a frozen Qwen2.5-7B-Instruct reader on 300 two-hop MuSiQue records as one counterfactual paragraph is repeated, after the true paragraphs verbatim (left) and reworded (centre), and before them (right). MinHash suppresses whichever of the true and false paragraphs arrives second. The ledger is the only arm whose recall falls in no condition.

Table 2: Frozen-reader results under syndicated misinformation, 300 two-hop MuSiQue records, Qwen2.5-7B-Instruct. Gold is gold-answer recall and misled is the rate of reporting the planted answer. *Truth kept* is the fraction of records whose true final paragraph survives the rule, *worst case* the lowest gold recall of the three conditions, and tokens the mean context length at sixteen verbatim copies.

	Verbatim, ×16		Reworded, ×16		False first, ×1			
Accounting rule	Gold	Misled	Gold	Misled	Gold	Truth kept	Worst case	Tokens
Lineage ledger (this work)	0.5133	0.0100	0.5267	0.0133	0.5967	1.0000	**0.5133**	464.5
Canonical hashing (Lee et al., 2022)	0.5133	0.0100	0.3933	0.1533	0.5967	1.0000	0.3933	464.5
MinHash (Broder, 1997)	0.5500	0.0033	0.5267	0.0167	0.2667	0.4767	0.2667	386.3
SimHash (Charikar, 2002; Manku et al., 2007)	0.5133	0.0100	0.4633	0.0967	0.5567	0.9533	0.4633	454.4
No defence	0.3967	0.1367	0.3933	0.1533	0.5967	1.0000	0.3933	2362.1

6.2 Surface similarity is not lineage

Table 2 shows why near-duplicate detection, the deployed defence, cannot substitute for lineage. Canonical hashing (Lee et al., 2022) removes byte-identical copies and nothing else, so once each copy carries its own outlet framing it behaves like no defence. The ledger leads it by $+0.1333$ in gold recall, interval $[+0.0833, +0.1833]$, and changes the misled rate by -0.1400. SimHash (Charikar, 2002) at the published radius (Manku et al., 2007) catches some reworded copies and trails by $+0.0633$, interval $[+0.0200, +0.1067]$.

MinHash (Broder, 1997) fails more consequentially. The counterfactual differs from the true paragraph only in the answer span, so MinHash judges them near-duplicates and keeps whichever arrives first. When the false one arrives first it retains the true paragraph in only 0.4767 of records, and from a single copy the ledger leads it by $+0.3300$, interval $[+0.2767, +0.3833]$, and by $+0.2000$ under reworded framing. A near-duplicate detector cannot separate a copy from a contradiction, because the two differ only in lineage. The ledger discards no lineage-distinct source, and its worst-case recall in Table 2 is the highest of any rule.

6.3 Corroboration is directional

Section 3.3 argues that a source should be credited only along the directions its observation constrains. The real-data test asks whether that predicts something a retrieval system must know before it answers, namely whether the five retrieved MuSiQue paragraphs contain every supporting fact. Table 3 reports the pre-registered comparison over $2{,}417$ questions, of which 0.403 have sufficient evidence.

Keeping requirements separate is what makes support informative. In the pre-registered contrast, the weakest requirement's support exceeds the same support averaged over requirements by $+0.1535$ in stratified AUROC, interval $[+0.1312, +0.1755]$. The averaged form sits at chance within ev-

Table 3: Predicting whether five retrieved MuSiQue paragraphs contain every supporting paragraph, 2,417 questions, AUROC within hop-count strata, the pre-registered primary measure. Rows fix how lineage-distinct sources combine within a requirement (Section 3.3), and columns how requirements combine. The first row is the pre-registered contrast. The second was registered as a baseline, and its directional cell is the strongest score. Hop count alone, with pooled AUROC 0.7018, is constant within a stratum. Appendix C reports pooled values and copies.

Within a requirement	Weakest requirement (directional)	Mean requirement (scalar)
Sum over sources, registered primary	0.6563	0.5028
Best single source	**0.7132**	0.6930
Sum of retrieval scores, no requirements	0.4671	
Best retrieval score, no requirements	0.5467	

Table 4: Learned two-hop MuSiQue memory at sixteenfold delivery, five seeds, content readout, top-1 accuracy over 555 candidate answers, where a uniform guess scores 0.0018. *Re-chunked* delivers each paragraph with fifteen windows cut from it, and *fragments* delivers sixteen windows and withholds the paragraph. Credit is relative to the lineage-distinct ceiling. Appendix D gives fourfold delivery, NLL and MRR.

Accounting rule	Clean	Exact	Re-chunked	Fragments	Credit, fragments
Lineage ledger (this work)	0.1141	0.1141	0.1141	0.0832	1.000
Canonical hashing (Lee et al., 2022)	0.1141	0.1141	0.0815	0.0768	16.000
MinHash (Broder, 1997)	0.1141	0.1141	0.0816	0.0851	3.895
SimHash (Charikar, 2002)	0.1141	0.1141	0.0791	0.0819	11.849
Embedding detection (Xiao et al., 2024)	0.1141	0.1141	0.0834	0.0823	5.588
No defence	0.1141	0.0823	0.0815	0.0768	16.000

ery hop count, because a question can have many relevant sources and still no support for one of its hops. When four copies of the top paragraph are appended, the lineage-aware directional score exceeds provenance-free scalar counting by $+0.0481$, interval $[+0.0294, +0.0673]$. Both pre-registered real-data predictions hold, and the retrieval scores that systems commonly report, which ignore requirements, fall well below every directional score.

The strongest score keeps each requirement's best single source, and it is also directional, exceeding its own average over requirements by $+0.0202$, interval $[+0.0023, +0.0379]$, in an exploratory contrast. Summed support trails it by -0.0569, interval $[-0.0747, -0.0400]$, because sufficiency is a coverage question. A requirement is met once one passage states it, so a second lineage-distinct passage cannot change the label. Summation measures corroboration, which this label does not reward, and it treats a related passage's partial projection as a partial measurement. The best-source rule in turn cannot register a second source at all, so where corroboration matters the additive join is needed, and the ledger is what lets it credit a second source while ignoring a second copy.

6.4 Conservation in a learned memory

Table 4 reports a learned two-hop MuSiQue memory in three regimes, scored by a content readout that does not depend on how credit is scaled.

Under exact redelivery the ledger keeps its clean accuracy of 0.1141, because a redelivered hop is dropped before it reaches the recurrent content path, while the provenance-free control processes every copy and falls to 0.0823. The ledger leads it by $+0.0318$, interval $[+0.0168, +0.0468]$, changes negative log-likelihood by -0.8799 nats and raises mean reciprocal rank by $+0.0411$, every interval excluding zero. Exact and near-duplicate matching equal the ledger here, since byte-identical copies are what they catch.

Re-chunking separates what lineage does for accuracy from what it does for credit. When a paragraph arrives together with windows cut from it, as in an index that holds a document and its chunks, the ledger keeps the complete rendering, since every window ranks below it, while ev-

ery other rule loses more than a quarter of its accuracy. The ledger leads no defence by $+0.0326$, interval $[+0.0147, +0.0505]$, and the strongest detector, embedding detection, by $+0.0307$, interval $[+0.0226, +0.0388]$. When only windows arrive, the ledger can hold only a fragment. Its accuracy then lands within noise of reading every fragment, as predicted before the run, a difference of $+0.0064$ with interval $[-0.0059, +0.0187]$, and of the strongest detector, MinHash, while it credits one source where no defence credits sixteen. Lineage therefore always bounds credit, and it also protects accuracy whenever a complete rendering of the source is among the copies.

6.5 Transfer and efficiency

The ledger attaches without modification to a message-passing backbone (Gilmer et al., 2017), a neural cellular automaton (Mordvintsev et al., 2020) and a Set Transformer (Lee et al., 2019), conserving on each while the provenance-free variants track the delivery count. It costs less than no defence, since suppressed arrivals never reach the content path, running $4{,}674$ messages in 0.037 seconds against 0.063. Under benign sixteenfold redelivery a frozen Qwen2.5-1.5B-Instruct reader's context stays at 318.9 tokens against 3710.0 with recall unchanged (Appendices E and F).

7 Limitations

Source identity is not source truth. Nothing here discharges the assumption of Section 3.1 that a supplied identifier corresponds to a distinct observation, and Proposition 3 shows content alone cannot. Eight machine paraphrases carrying fabricated identifiers inflate credited evidence by 2.673, which Appendix G records as a quantified boundary rather than as Sybil robustness.

The frozen-reader results have two boundaries. When the true paragraph arrives first, MinHash suppresses the counterfactual as a near-duplicate of it, and the ledger minus MinHash difference is -0.0367, interval $[-0.0667, -0.0067]$, the same order dependence that produces its deficit in the reverse order. When sixteen copies precede the truth, the ledger changes the misled rate by -0.0267, interval $[-0.0500, -0.0033]$, and gold recall by $+0.0233$, whose interval $[-0.0200, +0.0667]$ includes zero. Under benign duplication the arms are indistinguishable on quality, so the benefit there is efficiency.

On AVeriTeC no variant separates on the official retrieved-track metric, and a pre-registered selective-prediction criterion returned a negative decision, a macro-F1 gain of $+0.0361$ with interval $[-0.0058, +0.0811]$. The learned models carry roughly 0.5M trainable parameters over a frozen 109M encoder, so larger scales remain to be measured.

The mechanism inherits the lineage its identifiers carry and trusts the version a sender asserts. URL-level identifiers do not group a story syndicated across outlets (Appendix G), and a sender who reuses a true source's identifier with a higher refinement level can replace its content at unchanged credit. Adversarial settings therefore need versions authenticated by the issuer.

The directional evidence is real-data evidence only. In a synthetic learned model, directional credit did not beat a scalar whose scale is learned, so two of three registered synthetic predictions failed and no synthetic directional claim is made (Appendix C). The real-data test uses the dataset's question decomposition as its requirement directions, which stands in for a decomposer, and its strongest score was registered as a baseline rather than as the primary.

8 Conclusion

Repetition is not corroboration, and corroboration has a direction. A memory that credits lineage-distinct sources, each only along what it observed, shields a frozen reader from a falsehood copied sixteen times without the order dependence that makes near-duplicate detection discard the truth, tells when retrieved evidence leaves a requirement unsupported, and keeps a learned memory's accuracy under redelivery, with a finite-sample guarantee. The three findings share one mechanism, credit bounded by the lineage a message carries rather than by how the message looks, so redelivery, re-chunking and fragments are all suppressed by the same join. Authenticating version keys against their issuer, and supplying identifiers from shared-error signals where provenance metadata is absent, remain open.

AI USE STATEMENT

Generative models are used in this work as objects of study. Open models at pinned revisions serve as frozen readers, and Qwen2.5-0.5B generated the eight paraphrase candidates per claim that form the Sybil audit of Appendix G, with revisions and decoding settings given in Appendix H. That candidate corpus was frozen before any evaluation, and a candidate was admitted only after passing a pinned natural language inference check in both directions and a surface-similarity threshold. The authors reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses two public research datasets and involves no human subjects. AVeriTeC is used under its released licence and is not redistributed, and MuSiQue is used and redistributed under CC BY 4.0. The counterfactual paragraphs are generated mechanically from public records, are framed with generic outlet labels rather than real organisations, and are released only as evaluation artifacts. Section 7 reports an attack that fabricates source identity, because a defence whose failure mode is quantified is safer than one whose failure mode is undocumented. A mechanism that bounds confidence could be misread as a guarantee of correctness. The guarantee is that credited evidence does not grow under repetition of a stable lineage, and it says nothing about whether the underlying source is true.

REPRODUCIBILITY STATEMENT

Every number reported here is read from a committed raw output. The syndication, reader, directional, learned-memory and taxonomy numbers are written by the scripts listed in Appendix H, and the remaining numbers are transcribed from the committed reports named there. The theoretical claims are stated with their assumptions in Appendix A, and every proof appears there in full. Appendix H records the hardware, the pinned revisions, the seed counts, the parameter counts and the regeneration commands. The syndication and directional experiments were pre-registered before any output existed, and Appendices B and C report their predictions and decision rules. The MuSiQue corpus ships compressed in the repository with a digest manifest. AVeriTeC is not redistributed, and the official knowledge store is used in place of live web search. The full source, including the test suite that guards the structural properties asserted here, is available anonymously.[1]

REFERENCES

Pablo O. Arambel. Structured pedigree information for distributed fusion systems. In *Signal Processing, Sensor Fusion, and Target Recognition XVII, Proceedings of SPIE*, volume 6968, pp. 69680X, 2008. doi: 10.1117/12.777422.

Andrei Z. Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences*, pp. 21–29, 1997.

KC Chang, Way Kuo, Yaakov Bar-Shalom, Chee-Yee Chong, and Shozo Mori. Primex - prime-based graph encoding and extraction for information fusion. In *28th International Conference on Information Fusion (FUSION)*, pp. 1–8, 2025. doi: 10.23919/fusion65864.2025.11124187.

Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 380–388, 2002.

Edith Cohen and Haim Kaplan. Summarizing data using bottom-k sketches. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Principles of Distributed Computing (PODC)*, pp. 225–234, 2007. doi: 10.1145/1281100.1281133.

[1]Anonymised code and data: `https://github.com/anony123456787654321-eng/ICLR_evnca2212`. The repository contains the committed raw outputs, the analysis scripts that produce every table and figure, and the pre-registration document.

Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Integrating conflicting data: The role of source dependence. *Proceedings of the VLDB Endowment*, 2(1):550–561, 2009a. doi: 10.14778/1687627.1687690.

Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Truth discovery and copying detection in a dynamic world. *Proceedings of the VLDB Endowment*, 2(1):562–573, 2009b. doi: 10.14778/1687627.1687691.

Xin Luna Dong, Laure Berti-Equille, Yifan Hu, and Divesh Srivastava. Global detection of complex copying relationships between sources. *Proceedings of the VLDB Endowment*, 3(1-2):1358–1369, 2010. doi: 10.14778/1920841.1921008.

John R. Douceur. The Sybil attack. In *International Workshop on Peer-to-Peer Systems (IPTPS)*, pp. 251–260, 2002.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272, 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

Simon J. Julier and Jeffrey K. Uhlmann. A non-divergent estimation algorithm in the presence of unknown correlations. In *Proceedings of the 1997 American Control Conference*, volume 4, pp. 2369–2373. IEEE, 1997.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, 2022.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, 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 (NeurIPS)*, volume 33, pp. 9459–9474, 2020.

Zhongxuan Liu, Sicheng Zhou, and Hongzhi Wang. Record grouping controls evidence weight in language models. *arXiv preprint arXiv:2609.08698*, 2026.

Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. Detecting near-duplicates for web crawling. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*, pp. 141–150, 2007.

Alexander Mordvintsev, Ettore Randazzo, Eyvind Niklasson, and Michael Levin. Growing neural cellular automata. *Distill*, 5(2), 2020. doi: 10.23915/distill.00023.

Kevin P. Murphy, Yair Weiss, and Michael I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 467–475, 1999.

Atharv Naphade. Rational synthesizers or heuristic followers? analyzing LLMs in RAG-based question-answering. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 40293–40311, 2026.

Suman Nath, Phillip B. Gibbons, Srinivasan Seshan, and Zachary R. Anderson. Synopsis diffusion for robust aggregation in sensor networks. In *Proceedings of the 2nd International Conference on Embedded Networked Sensor Systems (SenSys)*, pp. 250–262, 2004.

Seonghyun Park, Narae Ryu, Gahee Kim, Dongyeop Woo, Se-Young Yun, and Sungsoo Ahn. Non-backtracking graph neural networks. *Transactions on Machine Learning Research*, 2024.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 30811–30849, 2024. doi: 10.52202/079017-0970.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. AVeriTeC: A dataset for real-world claim verification with evidence from the web. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, pp. 65128–65167, 2023. doi: 10.52202/075280-2842.

Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. The automated verification of textual claims (AVeriTeC) shared task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pp. 1–26, 2024. doi: 10.18653/v1/2024.fever-1.1.

Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.

Marc Shapiro, Nuno Preguiça, Carlos Baquero, and Marek Zawirski. Conflict-free replicated data types. In *International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS)*, volume 6976 of *Lecture Notes in Computer Science*, pp. 386–400, 2011. doi: 10.1007/978-3-642-24550-3_29.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL `https://aclanthology.org/2022.tacl-1.31/`.

Sikun Wang, Yixi Zhou, Lei Fan, and Fan Zhang. GraphEcho: Structural redundancy and evidence provenance in LLM graph agents. *arXiv preprint arXiv:2609.17695*, 2026.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust RAG against retrieval corruption. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 392–410, 2026. doi: 10.1109/SaTML68715.2026.00029.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, 2024. doi: 10.1145/3626772.3657878.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations (ICLR)*, 2024.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844, 2025.

A ASSUMPTIONS AND COMPLETE PROOFS

This appendix states the assumptions behind Theorem 1, gives its formal statement and proof, proves Corollary 1, states and proves Proposition 1 in full, and establishes the two propositions on calibration and on lineage used in the main text.

A.1 ASSUMPTIONS

Each assumption is stated where it is first used, and the scope of every statement below is limited to the assumptions it names.

Assumption 1 (Lipschitz readout)**.** *The decoder g maps a normalized ledger summary h and a mass W to logits, and satisfies $\|g(h', W') - g(h, W)\|_\infty \le L_h\|h' - h\|_2 + L_W|W' - W|$ with constants that do not depend on the delivery count.*

Assumption 2 (Stable lineage)**.** *Root identifiers are assigned by an oracle and are not under adversarial control. Proposition 3 shows why this cannot be replaced by an inference from content.*

Assumption 3 (Independent coordinated hashing)**.** *Root hashes $u_r = h(r)$ are independent and uniform on $[0, 1)$, and the payloads and weights are fixed given the root set, so the retained subset is independent of the values it retains. The implementation approximates this idealization with a seeded 64-bit digest.*

A.2 FORMAL STATEMENT OF THEOREM 1

Fix root payloads $a_r \in \mathbb{R}^p$ and weights $w_r > 0$ independently of the sketch hashes. Write $W = \sum_r w_r$ and $h = \sum_r a_r/W$. A coordinated bottom-k sample I with cardinality budget $k_h \ge k$ defines

$$\hat{h} = \frac{\sum_{r\in I} a_r}{\sum_{r\in I} w_r}, \qquad \hat{W} = \frac{\hat{n}}{|I|}\sum_{r\in I} w_r.$$

Let $q = \operatorname{softmax} g(h, W)$ and $\hat{q} = \operatorname{softmax} g(\hat{h}, \hat{W})$ under Assumption 1. For $\delta \in (0, 1)$ the finite-population radii r_h and r_W of Equation 5 are functions of $(n, k, k_h, \{a_r, w_r\}, \delta)$. When the denominator condition $e < 1$ holds, with probability at least $1 - \delta$, putting $A = L_h r_h + L_W r_W$ gives

$$\mathrm{KL}(q \,\|\, \hat{q}) \le A^2/2, \qquad |-\log \hat{q}_y + \log q_y| \le 2A \quad \text{for every label } y. \tag{3}$$

An exact-readout top-two logit margin greater than $2A$ guarantees the same predicted class. For $n \le k$ all radii are zero. Replaying held versions changes neither readout nor the bound.

A.3 PROOF OF THEOREM 1

Radii. Put $u_r = a_r - w_r h$, so that $\sum_r u_r = 0$, and define

$$c = \begin{cases} n(n-k)/[k(n-1)] & n > k, \\ 0 & n \le k, \end{cases}$$
$$V_u = c\sum_r \|u_r\|_2^2, \qquad V_D = c\sum_r (w_r - W/n)^2, \qquad V_M = \mathrm{Var}(\hat{W}). \tag{4}$$

The mass variance V_M is finite when cardinality is exact or $k_h \ge 2$. For a chosen failure probability define

$$e = \frac{\sqrt{3V_D/\delta}}{W}, \quad r_h = \frac{\sqrt{3V_u/\delta}}{W(1-e)}, \quad r_W = \sqrt{3V_M/\delta}, \qquad e < 1. \tag{5}$$

When $e \ge 1$ the bound is vacuous. The radii use the full population, so they are conditional analytic bounds rather than certificates computable from a sketch alone. Because e is a Chebyshev radius, the denominator condition constrains the retained fraction k/n, and the bound is informative under mild compression.

Sampling step. For $n > k$, bottom-k selects a uniform k-subset independently of the ordered hash values. With $\tilde{U} = (n/k)\sum_I u_r$ and $\tilde{D} = (n/k)\sum_I w_r$, the simple-random-sampling moments are

$$\mathbb{E}\tilde{U} = 0, \quad \mathbb{E}\|\tilde{U}\|_2^2 = V_u, \quad \mathbb{E}\tilde{D} = W, \quad \mathrm{Var}(\tilde{D}) = V_D.$$

Subset indicators satisfy $\mathbb{E}I_r = k/n$ and $\mathbb{E}I_rI_s = k(k-1)/[n(n-1)]$. Expanding the squared norm and using $\sum_r u_r = 0$ gives V_u, and centering w_r gives V_D. Markov's inequality on $\|\tilde{U}\|_2^2$ and Chebyshev's inequality on $\tilde{D}$ and $\hat{W}$, each with failure probability $\delta/3$, give simultaneously $\|\tilde{U}\|_2 \le \sqrt{3V_u/\delta}$, $\tilde{D} \ge W(1-e)$ and $|\hat{W} - W| \le r_W$. The union bound requires no independence between these events. Since $\hat{h} - h = \tilde{U}/\tilde{D}$, it follows that $\|\hat{h} - h\|_2 \le r_h$. The case $n \le k$ is exact.

Neural and categorical steps. Let $z = g(h, W)$ and $v = g(\hat{h}, \hat{W}) - z$. Assumption 1 gives $\|v\|_\infty \le A$ on the same event. For $q = \mathrm{softmax}\, z$, direct substitution gives

$$\mathrm{KL}(q \,\|\, \hat{q}) = \log \mathbb{E}_{j\sim q} e^{v_j} - \mathbb{E}_{j\sim q} v_j \le \frac{(\max_j v_j - \min_j v_j)^2}{8} \le \frac{A^2}{2}, \tag{6}$$

where the first inequality is Hoeffding's lemma (Hoeffding, 1963). For each label the loss change is $\log \mathbb{E}_q e^{v_j} - v_y$. Both terms lie between $\min_j v_j$ and $\max_j v_j$, so its absolute value is at most $2A$. A top-two margin greater than $2A$ cannot be reversed by coordinate changes of magnitude at most A, which proves all three claims.

Scope. A decoder of the form $g(h, W) = A_2 \tanh(A_1 h + b_1) + b_2 + vW$ admits $L_h \le \|A_2\|_{2\to\infty}\|A_1\|_2$ and $L_W = \|v\|_\infty$, so the interface admits nonlinear neural computation. Because $\hat{n}$ cancels from $\hat{h}$, extra hash slots improve mass estimation but cannot repair content error when $L_W = 0$. The bound compares two model distributions rather than either model with the true conditional, and it applies to a fixed population rather than to populations chosen adaptively in response to the hashes. ■

A.4 Proof of Corollary 1

Proof. Let the ledger L contain root r at version v. Delivering r again at any version v' replaces the entry with $\arg\max_\kappa\{v, v'\}$. If $v' = v$ the result is v by idempotence. If $\rho_{v'} < \rho_v$ then v is retained. If $\rho_{v'} = \rho_v$ the $-m$ coordinate ensures that a version claiming greater mass cannot displace the incumbent. In every such case the ledger is unchanged as a value, so the credited quantity $\rho_{ir}\Lambda_r$ in Equation 2 and every downstream readout are unchanged. The joins are commutative and associative, so the conclusion holds for any order and any number of such deliveries. If instead $\kappa(v') > \kappa(v)$, the domain of the ledger is unchanged because r was already present, so credited evidence is unchanged and only the representation moves. Credit therefore depends on the number R of roots in the domain and not on the number N of arrivals. □

A.5 Unbiasedness of the Horvitz-Thompson readout

Let root r carry the additive payload η_r, so the quantity to estimate is $S = \sum_r \eta_r$ over $n > k$ roots.

Proof. Under Assumption 3 the sketch retains the k roots with the smallest hashes, and $\theta = u_{(k+1)}$ is the $(k+1)$st smallest hash. The argument conditions on ranks rather than on θ, since given θ exactly k roots fall below it and inclusions are not independent (Cohen & Kaplan, 2007). For a fixed root r, let τ_r be the kth smallest hash among the other $n-1$ roots. Then r is retained if and only if $u_r < \tau_r$, and whenever it is retained $\theta = \tau_r$, because r is then one of the k smallest and the $(k+1)$st smallest overall is the kth smallest of the rest. The estimator therefore gives root r the weight $\mathbf{1}\{u_r < \tau_r\}/\tau_r$. Since τ_r depends only on the other hashes, it is independent of u_r, and $\mathbb{E}[\mathbf{1}\{u_r < \tau_r\}/\tau_r \mid \tau_r] = 1$. Summing over roots gives $\mathbb{E}[\hat{S}] = \sum_r \eta_r = S$. □

A.6 Formal statement and proof of Proposition 1

Let the ledger hold n roots, root r at a version whose credited precision is $M_r = \rho_{ir}\Lambda_r \succeq 0$, and let $\Lambda = \Lambda_{\text{prior}} + \sum_r M_r$ as in Equation 2. For a unit vector u the support is $s(u) = u^\top \Lambda u$.

For unit requirement directions $d_1, \dots, d_H$ the directional confidence is $\gamma = \min_h s(d_h)$, and its conservative form, which keeps the best single source per requirement, is $\gamma^\vee = \min_h \big(d_h^\top \Lambda_{\text{prior}} d_h + \max_r d_h^\top M_r d_h\big)$. The scalar surrogate is $\Lambda^{\text{sc}} = \Lambda_{\text{prior}} + \sum_r \frac{1}{d}\text{tr}(M_r) I$.

Statement. (i) Λ, γ and $\gamma^\vee$ are functions of the ledger state, so they are unchanged by any permutation of arrivals and by any redelivery of a held version, and they depend on the R held roots and not on the $N \geq R$ arrivals. (ii) If $M_r u = 0$ for every r, then $s(u) = u^\top \Lambda_{\text{prior}} u$, whereas $\text{tr}\,\Lambda^{\text{sc}} = \text{tr}\,\Lambda$ and $u^\top(\Lambda^{\text{sc}} - \Lambda_{\text{prior}})u = \frac{1}{d}\sum_r \text{tr}(M_r)$ for every unit u, which is positive unless every M_r vanishes. Under an isotropic prior the surrogate's weakest-requirement confidence is therefore the same for every set of requirements. (iii) Let I be a coordinated bottom-k sample of the $n > k$ roots under Assumption 3, with the cardinality n known, and put $\hat{\Lambda} = \Lambda_{\text{prior}} + \frac{n}{k}\sum_{r \in I} M_r$. Then $\mathbb{E}\hat{\Lambda} = \Lambda$, and with probability at least $1 - \delta$

$$\sup_{\|u\|_2 = 1} \big|\hat{s}(u) - s(u)\big| \leq \|\hat{\Lambda} - \Lambda\|_2 \leq r_\Lambda = \sqrt{V_\Lambda/\delta}, \qquad V_\Lambda = c \sum_r \|M_r - \bar{M}\|_F^2, \tag{7}$$

with c as in Equation 4 and $\bar{M} = \frac{1}{n}\sum_r M_r$, so that $|\hat{\gamma} - \gamma| \leq r_\Lambda$ on the same event. For rank-one credit $M_r = \rho_r e_r e_r^\top$ with $\|e_r\|_2 = 1$ and $\rho_r \leq 1$, $r_\Lambda \leq n\sqrt{(n-k)/(k(n-1)\delta)}$. For $n \leq k$ the sketch is exact. An estimated cardinality adds its own variance, as the term V_M does in Theorem 1. (iv) The conservative confidence computed on the sample satisfies $\hat{\gamma}^\vee \leq \gamma^\vee$ with probability one.

Proof. (i) By Corollary 1, a redelivery that does not beat the incumbent leaves the ledger unchanged as a value, and the joins are commutative and associative, so every arrival order and every number of such redeliveries reach one ledger state. Λ, γ and $\gamma^\vee$ read only the held versions. A strictly better version replaces M_r for a root already held and leaves the number of summands at R.

(ii) $s(u) - u^\top \Lambda_{\text{prior}} u = \sum_r u^\top M_r u = 0$ when every $M_r u = 0$. The trace is linear and $\text{tr}\big(\frac{1}{d}\text{tr}(M_r) I\big) = \text{tr}(M_r)$. For unit u, $u^\top\big(\frac{1}{d}\text{tr}(M_r) I\big)u = \frac{1}{d}\text{tr}(M_r)$, and a nonzero positive semidefinite matrix has positive trace. With $\Lambda_{\text{prior}} = \lambda I$ the surrogate gives every requirement the support $\lambda + \frac{1}{d}\sum_r \text{tr}(M_r)$.

(iii) Put $x_r = \text{vec}(M_r)$ and $y_r = x_r - \bar{x}$, so $\sum_r y_r = 0$ and, since $\frac{n}{k}\sum_{r \in I} \bar{x} = \sum_r x_r$, $\text{vec}(\hat{\Lambda} - \Lambda) = \frac{n}{k}\sum_{r \in I} y_r$. As in the proof of Theorem 1, I is a uniform k-subset independent of the payloads, with $\mathbb{E}I_r = k/n$ and $\mathbb{E}I_r I_s = k(k-1)/[n(n-1)]$. Hence $\mathbb{E}\hat{\Lambda} = \Lambda$, and using $\sum_{r \neq s} y_r^\top y_s = -\sum_r \|y_r\|_2^2$,

$$\mathbb{E}\Big\|\tfrac{n}{k}\textstyle\sum_{r \in I} y_r\Big\|_2^2 = \tfrac{n^2}{k^2}\Big(\tfrac{k}{n} - \tfrac{k(k-1)}{n(n-1)}\Big)\textstyle\sum_r \|y_r\|_2^2 = c\sum_r \|y_r\|_2^2 = V_\Lambda.$$

Markov's inequality on $\|\hat{\Lambda} - \Lambda\|_F^2$ gives $\|\hat{\Lambda} - \Lambda\|_F \leq r_\Lambda$ with probability at least $1 - \delta$. For a symmetric matrix A and unit u, $|u^\top A u| \leq \|A\|_2 \leq \|A\|_F$, and $|\min_h a_h - \min_h b_h| \leq \max_h |a_h - b_h|$ transfers the bound to γ. For rank-one credit, $\sum_r \|y_r\|_2^2 \leq \sum_r \|x_r\|_2^2 = \sum_r \rho_r^2 \leq n$, so $V_\Lambda \leq cn = n^2(n-k)/[k(n-1)]$.

(iv) For each h, a maximum over the retained roots cannot exceed the maximum over all roots, and taking the minimum over h preserves the inequality. □

Part (ii) makes precise the claim of Section 3.3 that a scalar count keeps the total information and discards its direction. Part (iii) is the directional analogue of the representation radius of Theorem 1, and like it the radius involves only the held roots.

A.7 The calibration penalty of over-counted precision

Proposition 2 (Calibration penalty)**.** *Let one underlying observation of precision Λ be delivered N times, and let a provenance-free readout credit every arrival, so that it reports precision $N\Lambda$ where the correct posterior carries Λ. For d-dimensional Gaussian posteriors sharing a mean, the divergence from the correct posterior to the reported one is*

$$\text{KL}(N) \;=\; \tfrac{d}{2}\big(N - 1 - \ln N\big), \tag{8}$$

which vanishes at $N = 1$ and grows without bound. A conserving readout reports Λ at every N, so its divergence under the same model is zero.

Proof. For Gaussians with a common mean and precisions Λ and $N\Lambda$, the divergence is

$$\tfrac{1}{2}\big(\mathrm{tr}(N\Lambda\Lambda^{-1}) - d - \ln\det(N\Lambda\Lambda^{-1})\big) = \tfrac{1}{2}(Nd - d - d\ln N),$$

which is Equation 8. Non-negativity and monotonicity on $N \geq 1$ follow because the derivative $\frac{d}{2}(1 - 1/N)$ is non-negative there. □

Equation 8 states the scale of the cost of over-counting in the idealised case the proposition names. A learned model violates its shared mean premise, because its predicted mean also moves with the evidence it is given, so the expression is used here as a statement of scale and not as a prediction of any measured quantity.

A.8 Content does not determine lineage

Proposition 3 (Indistinguishability)**.** *There exist two generative processes, one emitting m copies of a single observation and one emitting m separately acquired observations, that induce identical distributions over the visible content. No function of content alone distinguishes them, so any accounting rule that reads content only must either over-credit the first process or under-credit the second.*

Proof. Let a source emit a fixed value v almost surely. Process A draws once from one such source and copies the result m times. Process B draws once from each of m distinct such sources. Both produce the content multiset $\{v\}^m$ with probability one, so the induced distributions over visible content are equal and no measurable function of content separates them. Evidence conservation must credit A once and B m times, so a content-only rule that is correct on A is wrong on B and conversely. □

The syndication experiment exhibits the converse direction of the same fact. A counterfactual that edits one span of a true paragraph shares almost all of its content yet descends from a different source, and Section 6.2 shows what a content-only rule does with it.

B Syndication experiment in full

This appendix reproduces the pre-registered design, reports the outcome of every prediction, and tabulates every condition. The pre-registration document was committed to the repository before any reader call was made, and the verdict below is computed from the per-record outputs by a script that applies the committed decision rules mechanically.

Question and construction. Xie et al. (2024) report that language models follow the side with more evidence when the supporting pieces are distinct texts. The experiment asks whether a frozen reader treats copies of one false source as additional evidence, and whether reading from the ledger removes that exposure. A record is eligible only if no gold answer or alias survives in the counterfactual, the substitute appears in it, and the substitute appears in neither true paragraph. Substitutes match the kind of the gold answer, whether a year, another number or an entity of similar length, since an implausible substitute would be rejected on plausibility alone. Reworded copies are framed as "Outlet k reports" with a trailing syndication marker, so no two copies share bytes while all share one lineage.

Fixed in advance. The reader is Qwen2.5-7B-Instruct (Qwen et al., 2024) with greedy decoding and 48 new tokens. The records are the first 300 eligible in file order, and no record was added after any output was seen. Intervals use 10,000 bootstrap resamples over records with percentile bounds. A context-building pass with no reader loaded established that MinHash and SimHash can suppress the counterfactual or the true paragraph as near-duplicates of each other, which is why the order-dependence prediction was registered.

Table 5: Pre-registered predictions and their outcomes. Differences are paired over the 300 records. All 5 of 5 predictions hold.

Prediction	Contrast	Difference	95% interval
P1, the phenomenon	No defence, sixteen minus one verbatim copy, gold	-0.1167	$[-0.1600, -0.0733]$
P2, protection	Ledger minus no defence, sixteen verbatim copies, gold	$+0.1167$	$[+0.0733, +0.1600]$
P3, misled rate rises	No defence, sixteen minus one verbatim copy, misled	$+0.1267$	$[+0.0900, +0.1667]$
P4, hashing fails on rewording	Canonical hashing minus no defence, sixteen reworded copies, gold	$+0.0000$	$[+0.0000, +0.0000]$
P5, order dependence	Ledger minus MinHash, one copy before the truth, gold	$+0.3300$	$[+0.2767, +0.3833]$

Table 5 lists the outcomes, and Table 6 reports gold-answer recall in every condition. The ledger's verbatim rows stay constant across multiplicity by construction, since repeated copies of one root are no-ops. The difference between the ledger and no defence at sixteen copies therefore equals the amount by which repetition moves the reader.

Table 6: Gold-answer recall in every syndication condition, 300 records, Qwen2.5-7B-Instruct.

Accounting rule	Copies	$\times 1$	$\times 2$	$\times 4$	$\times 8$	$\times 16$
False copies after the truth						
Lineage ledger (this work)	verbatim	0.5133	0.5133	0.5133	0.5133	0.5133
Lineage ledger (this work)	reworded	0.5200	0.5333	0.5367	0.5400	0.5267
Canonical hashing (Lee et al., 2022)	verbatim	0.5133	0.5133	0.5133	0.5133	0.5133
Canonical hashing (Lee et al., 2022)	reworded	0.5200	0.4933	0.4800	0.4733	0.3933
MinHash (Broder, 1997)	verbatim	0.5500	0.5500	0.5500	0.5500	0.5500
MinHash (Broder, 1997)	reworded	0.5300	0.5333	0.5267	0.5300	0.5267
SimHash (Charikar, 2002; Manku et al., 2007)	verbatim	0.5133	0.5133	0.5133	0.5133	0.5133
SimHash (Charikar, 2002; Manku et al., 2007)	reworded	0.5233	0.5033	0.4900	0.4833	0.4633
No defence	verbatim	0.5133	0.4933	0.4867	0.4333	0.3967
No defence	reworded	0.5200	0.4933	0.4800	0.4733	0.3933
False copies before the truth						
Lineage ledger (this work)	verbatim	0.5967	0.5967	0.5967	0.5967	0.5967
Lineage ledger (this work)	reworded	0.5767	0.5667	0.5700	0.5700	0.5733
Canonical hashing (Lee et al., 2022)	verbatim	0.5967	0.5967	0.5967	0.5967	0.5967
Canonical hashing (Lee et al., 2022)	reworded	0.5767	0.4733	0.5200	0.5367	0.5433
MinHash (Broder, 1997)	verbatim	0.2667	0.2667	0.2667	0.2667	0.2667
MinHash (Broder, 1997)	reworded	0.3767	0.3667	0.3667	0.3700	0.3667
SimHash (Charikar, 2002; Manku et al., 2007)	verbatim	0.5567	0.5567	0.5567	0.5567	0.5567
SimHash (Charikar, 2002; Manku et al., 2007)	reworded	0.5533	0.4700	0.4900	0.5100	0.4967
No defence	verbatim	0.5967	0.5200	0.5767	0.5867	0.5733
No defence	reworded	0.5767	0.4733	0.5200	0.5367	0.5433

C DIRECTIONAL CREDIT EXPERIMENTS IN FULL

Both experiments follow a pre-registration committed before any output existed, together with two dated amendments that were also committed before any output. The first amendment appended the copies in the real-data duplicated condition to the retrieved set, because exact copies tie with their original under cosine retrieval and would otherwise fill every slot and make the label constant. The second followed an independent audit of the design. It added a scalar arm whose overall scale is learned, which became the primary synthetic comparator. It also stratified the primary real-data test by hop count and added hop count, best-single-source and retrieval-score baselines, because the number of decomposition steps equals the number of supporting paragraphs and hop count alone predicts sufficiency. `analysis/directional_analysis.py` applies the registered rules to the committed outputs.

Synthetic. The task is the evidence graph of Appendix F, with a four-dimensional latent, three roots, eight cells on a path, eight message steps, $4{,}000$ training iterations and five seeds. Sources observe one, two or four directions. Every arm shares one content path. Directional credit uses each root's precision, the scalar arms use $\frac{1}{d}\mathrm{tr}(\Lambda_r)I$, and the fitted scalar arm multiplies that by a learned global scale. Table 7 reports the negative log-likelihood of the true latent. Directional credit has the lowest error at every rank, but its registered advantage over the fitted scalar at rank one, $+0.1103$

nats with interval $[-0.1765, +0.3971]$, does not exclude zero, and the advantage grows rather than shrinks with rank, so the registered predictions S1 and S3 fail and no synthetic directional claim is made. The lineage prediction S2 holds, with $+21.8666$ nats, interval $[+4.9638, +38.7694]$.

Table 7: Synthetic directional credit, five seeds. Negative log-likelihood of the true latent under clean delivery at each observation rank, under sixteenfold duplication at rank one, and coverage of the marginal 95% interval at rank one. The pre-registration named 90% coverage as a secondary metric; the coverage routine the model already used measures 95%, and that is what is reported.

Credit	Rank 1	Rank 2	Rank 4	Rank 1, $\times 16$	95% coverage
Lineage, directional (this work)	5.4535	4.7639	3.8581	6.8171	0.9184
Lineage, scalar with learned scale	5.5639	4.9065	4.0863	6.7580	0.9094
Lineage, scalar	5.7140	5.1473	4.4186	7.1349	0.8875
No lineage, directional	5.4535	4.7639	3.8581	28.6837	0.9184
No lineage, scalar	5.7140	5.1473	4.4186	31.2987	0.8875

Real data. Paragraphs and queries are encoded with BGE `bge-base-en-v1.5` (Xiao et al., 2024) at a pinned revision, with the retrieval instruction applied to queries only. Intervals come from $10{,}000$ paired bootstrap resamples over records, resampled within hop strata for the stratified tests. Table 8 reports every score. The best-single-source scores use the cosine rather than its square. Every record's best cosine is at least 0.2747, so the two order records identically, and the score is the rank-one support of the single best source, the conservative form $\gamma^{\vee}$ of Appendix A. The contrast between its weakest and mean requirement was not registered and is reported as exploratory. Pooled over hop counts, the best-source score reaches 0.7404, above hop count alone at 0.7018.

Table 8: Predicting whether five retrieved MuSiQue paragraphs contain every supporting paragraph. AUROC pooled over records and stratified by hop count, and pooled AUROC with four copies of the top paragraph appended and counted per arrival. Counted once per lineage, the copies leave every score at its clean value by construction. Weakest-requirement rows are directional and mean-requirement rows scalar.

Score	Clean	Clean, stratified	Copies, per arrival
Weakest requirement, sum over sources (registered primary)	0.6989	0.6563	0.6874
Mean requirement, sum over sources	0.6019	0.5028	0.6508
Weakest requirement, best single source	0.7404	0.7132	0.7404
Mean requirement, best single source	0.7151	0.6930	0.7151
Sum of retrieval scores	0.5006	0.4671	0.5633
Maximum retrieval score	0.6137	0.5467	0.6137
Hop count alone	0.7018	0.5000	0.7018

D DUPLICATED-HOP CONSTRUCTION

The learned memory transports evidence along MuSiQue chains, one hop per delivery, and each regime rewrites the delivery stream of a chain while keeping the root identifier of every arrival. Every construction is deterministic, so intervals come from seeds of the learned model rather than from randomness in the stream. Table 9 reports every arm in every regime with the readout and seeds of Table 4.

Exact. An arrival is repeated verbatim.

Re-chunked. At multiplicity m the paragraph arrives once in full and then as $m - 1$ distinct contiguous windows cut from it, with widths nearest seventy percent of the paragraph and offsets spread across it, so no window adds a fact and every window has its own surface hash. The windows at a lower multiplicity are a prefix of those at a higher one. A window carries a subset of its paragraph's content, so it ranks below the paragraph in the version key by the number of words it omits. The ledger therefore holds the whole paragraph and its content path sees exactly the clean stream,

while credit, which counts identifiers, is unaffected. Exact hashing admits every window as a new source, the near-duplicate detectors admit those they do not judge similar to an earlier arrival, and the provenance-free control processes all of them. The detectors suppress some windows, as their credit in Table 10 shows, but the windows they admit move the content path about as far as all sixteen renderings move the undefended memory.

Fragments. The same windows arrive without the paragraph, m of them at multiplicity m, so no complete rendering exists and every rule, the ledger included, can hold only a fragment. Two predictions were committed before any output of this regime existed, namely credit of exactly one unit per source and an accuracy difference against no defence whose interval includes zero at sixteenfold, and both hold. A window that outranks the one held replaces it and reaches the content path, so at sixteenfold 7.5 of 32 arrivals per record reach the ledger's content path, which is why its accuracy is lower at sixteenfold than at fourfold. An earlier fragments-only run with at most three distinct windows gave the same result, a ledger minus no defence difference of -0.0040 with interval $[-0.0135, +0.0055]$.

Table 9: Learned two-hop MuSiQue memory in every duplication regime, five seeds, content read-out. Accuracy is top-1 over 555 candidate answers, and negative log-likelihood and mean reciprocal rank are at sixteenfold delivery.

Accounting rule	Clean	$\times 4$	$\times 16$	NLL, $\times 16$	MRR, $\times 16$
Exact					
Lineage ledger (this work)	0.1141	0.1141	0.1141	5.5337	0.1986
Canonical hashing (Lee et al., 2022)	0.1141	0.1141	0.1141	5.5337	0.1986
MinHash (Broder, 1997)	0.1141	0.1141	0.1141	5.5337	0.1986
SimHash (Charikar, 2002)	0.1141	0.1141	0.1141	5.5337	0.1986
Embedding detection (Xiao et al., 2024)	0.1141	0.1141	0.1141	5.5337	0.1986
No defence	0.1141	0.0895	0.0823	6.4136	0.1575
Re-chunked					
Lineage ledger (this work)	0.1141	0.1141	0.1141	5.5337	0.1987
Canonical hashing (Lee et al., 2022)	0.1141	0.0923	0.0815	6.7219	0.1501
MinHash (Broder, 1997)	0.1141	0.0923	0.0816	6.6558	0.1535
SimHash (Charikar, 2002)	0.1141	0.0923	0.0791	6.6715	0.1497
Embedding detection (Xiao et al., 2024)	0.1141	0.0877	0.0834	6.7016	0.1541
No defence	0.1141	0.0923	0.0815	6.7219	0.1501
Fragments					
Lineage ledger (this work)	0.1141	0.1050	0.0832	6.5506	0.1593
Canonical hashing (Lee et al., 2022)	0.1141	0.0957	0.0768	6.7492	0.1479
MinHash (Broder, 1997)	0.1141	0.0896	0.0851	6.6227	0.1573
SimHash (Charikar, 2002)	0.1141	0.0955	0.0819	6.7004	0.1536
Embedding detection (Xiao et al., 2024)	0.1141	0.0896	0.0823	6.7000	0.1534
No defence	0.1141	0.0957	0.0768	6.7492	0.1479

E FUSION ARMS AS ACCOUNTING RULES

The Dempster and covariance intersection arms of Table 10 share the content path of the provenance-free control and differ only in how they assign credit, which makes the comparison one of accounting alone. Dempster's rule (Shafer, 1976) treats each arrival as a simple support function with a frozen mass of $m = 0.5$. Combining n such functions gives belief $1 - (1 - m)^n$, which in units of one delivery is a credit of $(1 - (1 - m)^n)/m$. At sixteen deliveries this is 1.99997, and it approaches $1/m = 2$ as n grows. Covariance intersection (Julier & Uhlmann, 1997) assigns each root a convex weight, so a root receives the fraction of all arrivals that it contributed. Under uniform duplication of R roots each receives $1/R$. Because the three arms share one content path, their task predictions are identical in every condition, and the repository verifies that identity directly. The best-single-source score of Section 6.3 is covariance intersection along each requirement and shares its profile. It is idempotent, so copies cannot inflate it, and it is not additive, so a second lineage-distinct source cannot raise it.

Table 10: Accounting properties and credited evidence under re-chunking at sixteenfold duplication on two-hop chains (Appendix D), where A1 to A4 are order invariance, idempotence within a root, additivity across roots and bounded state. Credit is relative to the lineage-distinct ceiling, so 1.000 is exact conservation. The origin column records where each value comes from. Values marked *definition* or *algebra* follow from the rule itself, and the implementation reproduces them exactly, which verifies the implementation rather than testing a hypothesis. Values marked *measured* depend on how the duplicates were constructed and admit no closed form.

Accounting rule	A1	A2	A3	A4	Credit	Origin
Lineage ledger (this work)	✓	✓	✓	✓	**1.000**	definition
No defence	✓	–	✓	–	16.000	definition
Dempster's rule (Shafer, 1976)	✓	–	✓	✓	2.000	algebra, $1/m$
Covariance intersection (Julier & Uhlmann, 1997)	✓	✓	–	✓	0.500	algebra, $1/R$
Surface deduplication (Lee et al., 2022; Broder, 1997; Charikar, 2002; Xiao et al., 2024)	✓	–	✓	✓	4.823 to 16.000	measured
Non-backtracking (Park et al., 2024)	✓	–	✓	–	16.000	measured

F TRANSFER ACROSS BACKBONES AND COST

Table 11 reports credited evidence against delivery multiplicity for every arm of the synthetic graph matrix, and Figure 2 plots the same measurements. The message-passing backbone (Gilmer et al., 2017) conserves exactly, with a seed standard deviation of 0.0000 at every multiplicity. The cellular automaton (Mordvintsev et al., 2020) conserves to within 0.0142 standard deviation at sixteenfold delivery, with credit between 0.9730 and 1.0113 across seeds. The residual comes from stochastic asynchronous firing, which means a given duplicate may or may not reach a given cell within the step budget. The same mechanism explains why its provenance-free arm reaches 14.2792 rather than 16.0000. Non-backtracking message passing (Park et al., 2024) removes the immediate reverse message, which addresses short cyclic feedback, and it inflates exactly as no defence does.

Table 11: Credited evidence against delivery multiplicity on the synthetic graph task, five seeds. Credit is relative to the single-source baseline, so a conserving arm stays near 1.000 and a provenance-free arm tracks the multiplicity.

Arm	1×	2×	4×	8×	16×
Message passing (Gilmer et al., 2017), conserving	1.0000	1.0000	1.0000	1.0000	**1.0000**
Cellular automaton (Mordvintsev et al., 2020), conserving	1.0000	1.0072	0.9967	0.9952	**0.9932**
Non-backtracking (Park et al., 2024)	1.0000	2.0000	4.0000	8.0000	16.0000
Cellular automaton (Mordvintsev et al., 2020), provenance-free	1.0000	1.9179	3.5904	7.7245	14.2792
Message passing (Gilmer et al., 2017), provenance-free	1.0000	2.0000	4.0000	8.0000	16.0000

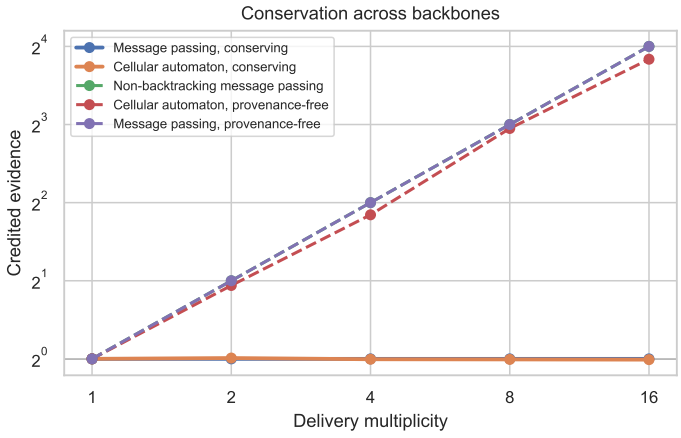

Figure 2: Credited evidence against delivery multiplicity on log-log axes, five seeds. The horizontal rule marks exact conservation. Conserving arms stay on that rule, while provenance-free arms and the non-backtracking baseline track the identity.

Duplication also degrades a matched provenance-free message-passing network by 6.144 nats more in negative log-likelihood than the conserving one. Table 12 reports end-to-end time on a common workload. The ledger suppresses arrivals before they reach the content path, which is why its total falls below that of no defence. MinHash and embedding detection spend roughly eight times as long as the ledger in their accounting stage.

Table 12: End-to-end cost on a common workload of $4{,}674$ delivered messages. Ledger time is the accounting overhead and content time is the learned path.

Accounting rule	Delivered	Reaching content	Accounting (ms)	Content (s)	Total (s)
Lineage ledger (this work)	4,674	984	1.36	0.030	**0.037**
Canonical hashing (Lee et al., 2022)	4,674	984	1.34	0.031	0.037
SimHash (Charikar, 2002)	4,674	984	1.42	0.031	0.038
MinHash (Broder, 1997)	4,674	984	10.65	0.031	0.080
Embedding detection (Xiao et al., 2024)	4,674	984	11.15	0.030	0.082
No defence	4,674	4,674	0.07	0.063	0.063

The frozen-reader efficiency measurement of Section 6.5 uses Qwen2.5-1.5B-Instruct (Qwen et al., 2024) on the first 120 two-hop records.

G SEMANTIC-SYBIL AUDIT

Fabricated identities can inflate credited evidence, and the audit measures how far they do when each fabricated source carries a meaning-preserving paraphrase.

Generation and acceptance. For each of 200 AVeriTeC claims, eight paraphrase candidates of the source are generated with Qwen2.5-0.5B (Qwen et al., 2024) at a pinned revision under deterministic decoding. The candidate corpus is frozen before any evaluation. A candidate is accepted only if it entails and is entailed by the original under a pinned natural language inference model, and only if its surface similarity to the original is below 0.87. The first test excludes candidates that changed the meaning, and the second excludes near-copies that surface deduplication would already catch.

Semantic grouping. The BGE (Xiao et al., 2024) cosine threshold of 0.88936 is fitted on training pairs only, under a maximum false-link budget of 0.5% fixed in advance. On $15,030$ positive and $9,233$ negative training pairs it attains recall 0.9354 at a realised false-link rate of 0.50%. Of 200 claims, 60 retain enough accepted paraphrases to support multiplicity eight.

Table 13: Credited evidence under the paraphrase attack, five seeds over 60 eligible claims. Credit is relative to the single-source baseline.

Grouping	Multiplicity	Credit ratio	95% interval	NLL	Flip rate
URL identity	1	1.000	[1.000, 1.000]	1.1212	0.0000
	2	2.060	[1.995, 2.126]	1.1056	0.0467
	4	4.131	[4.030, 4.233]	1.1496	0.0433
	8	4.567	[4.328, 4.807]	1.1394	0.0600
canonical URL	1	1.000	[1.000, 1.000]	1.1212	0.0000
	2	2.060	[1.995, 2.126]	1.1056	0.0467
	4	4.131	[4.030, 4.233]	1.1496	0.0433
	8	4.567	[4.328, 4.807]	1.1394	0.0600
surface family	1	1.000	[1.000, 1.000]	1.1212	0.0000
	2	2.016	[1.945, 2.088]	1.1046	0.0467
	4	3.363	[3.230, 3.497]	1.1302	0.0433
	8	3.724	[3.514, 3.933]	1.1198	0.0600
semantic family	1	1.000	[1.000, 1.000]	1.1212	0.0000
	2	1.676	[1.598, 1.754]	1.0944	0.0467
	4	2.522	[2.362, 2.682]	1.1090	0.0433
	8	2.673	[2.468, 2.878]	1.0968	0.0600

Table 13 shows that semantic grouping reduces the attack relative to identifier-based grouping, from 4.567 to 2.673 at multiplicity eight, and that no grouping restores exact conservation. The negative log-likelihood moves little across multiplicities, so fabricated identities inflate credit more than they move the prediction.

H REPRODUCIBILITY AND COMPUTE

Models and revisions. Passages for the learned models are encoded with the 768-dimensional BGE model (Xiao et al., 2024), released as `bge-base-en-v1.5`, at revision `a5beb1e3`. The frozen readers are Qwen2.5-7B-Instruct and Qwen2.5-1.5B-Instruct (Qwen et al., 2024) in bfloat16 with greedy decoding and 48 new tokens, and the paraphrase generator is Qwen2.5-0.5B. No encoder, reader or generator is fine-tuned anywhere in this work.

Hardware. The synthetic backbone matrix and the synthetic directional models run on CPU. The learned MuSiQue memory, the Sybil audit and the frozen readers run on an NVIDIA H100 80GB partition with PyTorch 2.5.1.

Seeds and scale. The learned models use five seeds, 10,000 iterations and a batch size of 128, with roughly 0.5M trainable parameters over the frozen encoder. The frozen readers are deterministic, so their uncertainty is estimated by resampling records.

Regeneration. Each script below reads committed raw outputs and writes the named artifacts.

- `experiments/musique_syndication_eval.py`, run once per arrival order, produces the frozen-reader syndication outputs.
- `analysis/syndication_analysis.py` applies the pre-registered decision rules and writes the verdict.
- `analysis/story_tables.py` writes Tables 2, 5 and 6, Figure 1, and every syndication and reader number quoted in the text.
- `experiments/musique_duplication_eval.py` and `experiments/musique_reader_eval.py` produce the learned-memory and reader-efficiency outputs, and `scripts/dgx_rechunk.sh` runs the re-chunked and fragments sweeps.
- `analysis/learned_memory_table.py` produces Tables 4 and 9.
- `experiments/directional_synthetic.py`, `experiments/directional_musique.py` and `analysis/directional_analysis.py` produce the directional results and their verdict.
- `analysis/axiom_predictions.py` and `analysis/conservation_figures.py` produce Table 10 and Figure 2.