ORIGINCOUNT: COUNTING ORIGIN EVENTS INSTEAD OF PAGES IN RETRIEVAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Many retrieval agents decide how sure they are, and when to stop searching, by counting the retrieved pages that support an answer. This count treats every page as a separate observation. On the web, however, many pages repeat a fact first reported in one place, the origin of these pages. On six web snapshots with provenance labels, about four in five retrieved pages are copies of an origin that another retrieved page reports, and counting the origins that provenance reveals instead of pages changes the stopping round or the answer on 18% of the questions on average. When such a copy depends on the answer only through its origin, the page count adds the evidence of that origin once for every copy, and the agent becomes confident too early. We propose OriginCount, which groups the retrieved pages by provenance signals (citations, shared identifiers, and restated content), gives each group one vote, and keeps each untraced page as its own group. We prove that among counts that add up page support with equal weight inside a group and count a lone page as one vote, one vote per group is the only one that copies do not change, and that with correct links the count lies between the true origin count and the page count. On a Qwen3-32B agent at a fixed budget, OriginCount lowers the rate of confidently wrong answers from 16.4% to 11.3% and the expected calibration error from 14.9% to 9.1%, with gains on five backbones, three retrievers, and five variants of the agent, at the cost of about one point of accuracy on average.

1 INTRODUCTION

Many retrieval agents answer questions by searching for pages, reading them, and deciding when they have read enough (Lewis et al., 2020; Nakano et al., 2021; Yao et al., 2023; Jiang et al., 2023; Asai et al., 2024). They often turn the support for a candidate answer into a confidence and stop once it is high. The support is usually the number of retrieved pages that state the answer, sometimes weighted by how reliable each site is (Hwang et al., 2025). The count is the right statistic when every page is a separate observation of equal weight: each supporting page then multiplies the odds of the answer by the same factor, and a threshold on the count is a threshold on the posterior, as in a sequential test (Wald, 1945). An agent that counts pages thus assumes that retrieved pages are separate observations.

On the web they often are not. A revenue figure or a clinical trial result is first reported in one place and then repeated by many other pages (Leskovec et al., 2009). We call the record where a fact was first reported the *origin* of the pages that report it, and a retrieved page that repeats an origin already reported by another retrieved page a *copy*. On six web snapshots with provenance labels, about four in five retrieved pages are copies (Fig. 1c). A copy that depends on the answer only through its origin leaves the posterior unchanged, while the page count adds the evidence of the origin once more for every copy (Lemma 1). The agent then becomes confident too early, and a wrong origin with many copies can outvote correct origins with fewer copies. Counting the origins that provenance reveals instead of pages changes the stopping round or the final answer on 16.5% to 20.2% of the questions.

Agents already use two tools that might remove such copies. De-duplication drops a page that is similar to one already kept (Broder, 1997; Abbas et al., 2023). It removes verbatim copies, but semantic de-duplication keeps 14% to 35% of rewritten copies (Fig. 1b). This is a limit of any similarity score: a rewritten copy can be as dissimilar as an independent report, and then no similarity threshold separates the two (Proposition 1). Source-reliability weighting asks how often a site is right (Hwang et al., 2025) and ignores whether a page is independent, so a copy from a trusted site gets

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.

more weight than a copy from a less reliable one (Fig. 1a). Whether two pages report one observation depends on where the content of each page came from, that is, its provenance (Buneman et al., 2001).

Provenance is often visible on the page itself. A page may cite the page it copies, carry the wire tag or dataset identifier of another page, or restate its content. OriginCount links two retrieved pages when one of these three signals applies, takes the groups of linked pages as origins, and gives each group one vote, the mean support of its pages. A page without links keeps its own vote, and every page stays in the context. The rule replaces only the counting step, so it fits any agent that counts support. Two formal results support this design. Among counts that add up page support with equal weight inside a group, one vote per group is the only one that copies do not change and that still counts a new origin as one vote (Theorem 1). When every link is correct, the count lies between the true origin count and the page count and equals the page count when nothing can be traced; when all pages of each origin agree, it removes, in expectation, the linked share of the double counting (Theorem 2).

On a Qwen3-32B agent at a fixed retrieval budget, OriginCount lowers the confident-error rate (CE), the share of questions answered wrongly with confidence at least 0.8, from 16.43% to 11.34%, and the expected calibration error from 14.93% to 9.08%. Its CE is the lowest of all compared methods except an oracle that groups pages by the true provenance labels (Table 1). The gain holds on all five backbones, three retrievers, and three types of evidence, and it adds to the gains of five agent variants (Table 2). It costs about one point of accuracy, a loss that is near zero on average with the true provenance labels, and an extra cost of 16 ms per page, or 2 ms with a rule-based extractor. This work measures how often retrieval agents count copies and why the common remedies keep them, characterizes the count that copies do not change, and turns it into a drop-in rule that we compare against 17 baselines and an oracle, with ablations that test whether each design choice is necessary.

2 Related Work

Retrieval agents and when to stop. Retrieval-augmented generation conditions a language model on retrieved passages (Lewis et al., 2020; Karpukhin et al., 2020). Agents interleave reasoning with search (Nakano et al., 2021; Yao et al., 2023; Trivedi et al., 2023; Press et al., 2023; Li et al., 2025; Jin et al., 2025) and decide when to retrieve or stop: FLARE retrieves when confidence drops (Jiang et al., 2023), Self-RAG learns reflection tokens (Asai et al., 2024), CRAG re-retrieves when evidence looks poor (Yan et al., 2024), adaptive methods route by question difficulty (Jeong et al., 2024; Su et al., 2024; Mallen et al., 2023), and Sufficient Context asks whether the context is enough (Joren et al., 2025). Like answer voting (Wang et al., 2023), they accumulate support before they answer. We compare against five of them in Table 1 and add OriginCount to five agent variants in Section 5.3.

De-duplication and redundancy. Near-duplicate detection by shingling and hashing is standard in web crawling (Broder, 1997; Broder et al., 1997; Charikar, 2002; Manku et al., 2007), and de-duplication improves language-model training (Lee et al., 2022), also with semantic embeddings of the text (Abbas et al., 2023). Diversity-aware reranking (Carbonell & Goldstein, 1998) and set selection (Lee et al., 2025) reduce the redundancy among retrieved passages, whose composition affects the answers (Cuconasu et al., 2024). These methods decide whether two pages are similar. Here we ask whether two pages are separate observations of a fact, which similarity alone cannot settle once heavily rewritten copies look much like independent reports of the same fact (see Proposition 1).

Dependent sources. Counting each cluster of dependent items once is standard practice in several fields: meta-analyses count a trial once however often it is published (Tramèr et al., 1997), coevolution methods weight each sequence by the inverse size of its cluster (Morcos et al., 2011), and treating dependent samples as independent is known as pseudoreplication (Hurlbert, 1984). Shared evidence is also a known source of double counting in data fusion (Julier & Uhlmann, 1997) and in inference on graphs with loops (Pearl, 1988). Truth discovery estimates the reliability of sources from their conflicting claims (Yin et al., 2008; Li et al., 2016; Dong et al., 2015); Dong et al. (2009a;b; 2010) showed that sources copy one another and that copied votes should be discounted, and Pochampally et al. (2014) modeled correlated data sources. These works detect copying from many shared claims across many data items, whereas a retrieval agent sees only about twenty pages per question, so we read dependence from the provenance signals that the pages themselves carry (Buneman et al., 2001).

Calibration, attribution, and correlated votes. We measure calibration by the expected calibration error (Naeini et al., 2015; Guo et al., 2017) and confident errors as in selective prediction (Geifman &

El-Yaniv, 2017; Kamath et al., 2020). Language models can state their own uncertainty with some accuracy (Kadavath et al., 2022; Tian et al., 2023), which also motivates the self-reported count among our baselines. Attribution methods check whether a statement is supported by a cited passage (Gao et al., 2023; Bohnet et al., 2022; Yue et al., 2023; Min et al., 2023), and work on knowledge conflicts studies passages that disagree (Xu et al., 2024; Xie et al., 2024); in our problem, passages agree because they share an origin. When distinct origins are correlated, the jury theorem (Boland, 1989; Berg, 1993) and the design effect (Kish, 1965) bound what any vote over such origins can gain.

3 Double Counting in Retrieval Agents

3.1 Problem Setup

A question x has a finite set of candidate answers. For a candidate a, let H_a be the event that a is correct. The agent retrieves pages in rounds, and after a round it has read pages $d_1, \dots, d_n$. A reader, the backbone language model, extracts from each page a support value $s_a(d_i) \in \{0, 1\}$ that is 1 when the page states a. The page count of a is

$$N_{\text{page}}(a) = \sum_{i=1}^{n} s_a(d_i). \tag{1}$$

The agent keeps a support count $N(a)$ for every candidate, which is $N_{\text{page}}(a)$ for the page-count agent. It answers the candidate $\hat{a}$ with the largest count, reports the confidence $c = f(N(\hat{a}))$ for an increasing map f, and stops at the first round at which $c \geq \tau$. The default agent uses $f(N) = \sigma(\beta N)$ with the logistic function σ, $\beta = 0.35$, and $\tau = 0.9$. An answer is a confident error when it is wrong and its reported confidence is at least 0.8, and CE is the share of all questions with a confident error.

Every page reports an observation that was first recorded somewhere, such as a wire report or a filing; this record is the origin $o(d)$ of page d. Pages with the same origin form a group, $\pi^\star$ is the partition of the read pages into groups, m_g is the size of group g, and G is the number of groups. In each group, every page is a copy except one: the origin record if it was read, and otherwise the first page read. Copies may repeat their origin verbatim or rewrite it. If the pages were independent given H_a and given its complement $\neg H_a$, every page that states a multiplied the odds of H_a by e^β, and every other page left the odds unchanged, then with prior odds of one, $f(N_{\text{page}})$ would equal the posterior.

3.2 Effect of Copies

Assumption 1 (Copies depend on the answer through their origin)**.** Condition on $\pi^\star$. For each group g, let E_g be the origin record of g and D_g the set of its copies. (i) $\mathbb{P}(D_g \mid E_g, H_a) = \mathbb{P}(D_g \mid E_g)$; (ii) the pairs (E_g, D_g) of distinct groups are independent given H_a and given $\neg H_a$.

Lemma 1 (Double counting)**.** *Under Assumption 1, if all origin records are read, the posterior log-odds of H_a are $L_0 + \sum_g \ell_g$, where L_0 is the prior log-odds given $\pi^\star$ and $\ell_g = \log \frac{\mathbb{P}(E_g \mid H_a)}{\mathbb{P}(E_g \mid \neg H_a)}$. A count that credits each of the m_g pages of group g with ℓ_g differs from the posterior log-odds by*

$$\Gamma = \sum_g (m_g - 1)\, \ell_g. \tag{2}$$

If every page of group g has support $s_{a,g}$ and $\ell_g = \beta s_{a,g}$, the posterior log-odds are $L_0 + \beta N_{\text{orig}}(a)$ with the origin count $N_{\text{orig}}(a) = \sum_g s_{a,g}$, and $\Gamma = \beta\,(N_{\text{page}}(a) - N_{\text{orig}}(a))$. If supporting origins arrive one per round, each with m supporting pages, the page count reaches the threshold τ after $\lceil N_\tau / m \rceil$ origins and the origin count only after $\lceil N_\tau \rceil$, where $N_\tau = f^{-1}(\tau)$.

The lemma is the known double counting of shared evidence (Pearl, 1988), written for retrieved pages; Appendix B.1 proves it and treats origin records that are not read. A copied origin inflates its own evidence m_g-fold, and the excess is larger for a more reliable origin. On a controlled set of pages that holds one origin (Fig. 1a), each added copy raises the confidence under the page count as much as a new origin would: four copies move the stopping round 0.53 rounds earlier, the same as four new origins, while under OriginCount the four copies leave both of them unchanged (Appendix E.2).

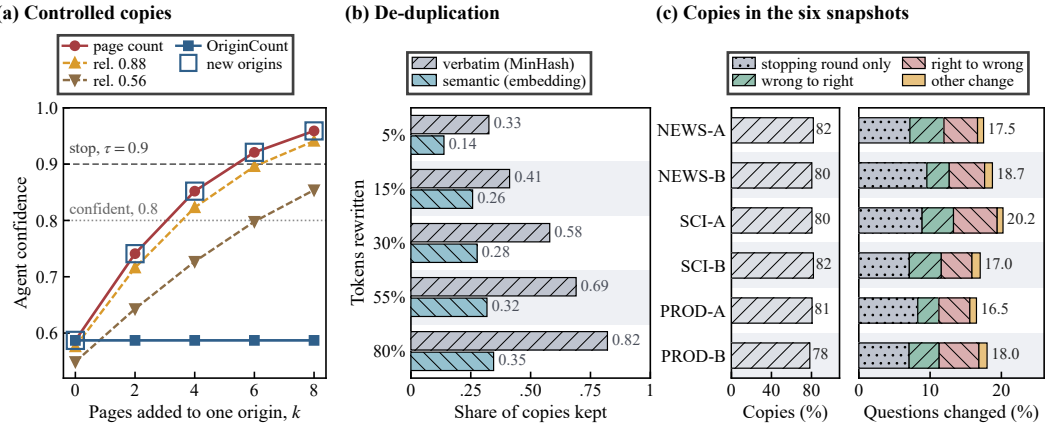


Figure 1: Copies inflate the page count, can pass de-duplication, and are common. (a) Confidence of the agent when k pages are added to a set of pages that holds one origin. The page count (red circles) and reliability weighting with site weight 0.88 or 0.56 (rel.) rise with copies, while OriginCount (blue squares) stays at $\sigma(0.35) = 0.587$; hollow squares: OriginCount when the added pages are new origins. Grey dashed line: stopping threshold τ; grey dotted line: confidence 0.8, the cutoff for a confident error. (b) Share of rewritten copies of one origin that survive verbatim (MinHash) and semantic (embedding) de-duplication, by the share of tokens rewritten. (c) Left: share of retrieved pages that are copies. Right: share of questions whose stopping round or answer changes when the agent counts origins with OriginCount, by kind of change; "other change" swaps one wrong answer for another. SCI and PROD abbreviate SCIENCE and PRODUCT.

Copies are common. On all six snapshots, 78% to 82% of the retrieved pages are copies (Fig. 1c), and questions see 4 to 5 pages per origin on average (Appendix E.4). Counting origins with OriginCount instead of pages changes the stopping round or the answer of 16.5% to 20.2% of the questions, and 39% to 51% of these changes only move the stopping round. The answer changes go both ways, since removing copies can take support from a wrong answer or from a right one (Appendix D.9). They partly cancel, so the net effect of copies lies mostly on the confidence and on when the agent stops.

3.3 De-duplication and Reliability Weighting

De-duplication removes pages close to one already kept. Verbatim de-duplication handles exact copies well: copying every page up to sixteen times moves CE by at most 0.16 points (Appendix E.4). Rewritten copies are much harder to remove. As rewriting grows from 5% to 80% of the tokens, verbatim matching keeps 33% to 82% of the copies and semantic matching 14% to 35% (Fig. 1b), and after semantic de-duplication CE reaches 96% to 99% of its full rise at a 30% rewrite. Proposition 1 shows why no threshold closes this gap once rewriting makes copies score like independent reports.

> **Proposition 1** (Error floor of merging by similarity)**.** *Let Z be any score, learned or not, computed from a pair of pages, with distribution F_{copy} over pairs of copies of one origin and F_{ind} over pairs of independent reports of the same answer. Every rule that decides from Z alone whether to merge a pair, a simple threshold included, has a false-split rate* FS *on copy pairs and a false-merge rate* FM *on independent pairs with* $\text{FS} + \text{FM} \geq 1 - \text{TV}(F_{\text{copy}}, F_{\text{ind}})$*, where* TV *is the total variation distance. In particular, one of the two error rates is at least* $(1 - \text{TV})/2$.

The bound is the two-point testing bound of Le Cam (1973) (see also Tsybakov, 2009) applied to the decision to merge (Appendix B.2). It holds for any signal computed from a pair of pages; signals differ only in TV. Rewriting moves the similarity of copies towards that of independent reports and shrinks TV for embedding similarity, so every threshold then keeps many copies, which leaves the double counting in place, or merges many independent reports, which deletes evidence. A citation or a shared identifier does not change when the text is reworded. Reliability weighting fails for another reason: it scales each page by the reliability of its site, so eight copies raise the weighted confidence from 0.576 to 0.941 at site weight 0.88 and from 0.549 to 0.854 at weight 0.56 (Fig. 1a). The number of copies also varies from one question to the next, so one global rescaling of f cannot undo this. Section 4 therefore reads the provenance of each retrieved page from the signals on the page itself.

4 ORIGINCOUNT

4.1 COUNTING RULE

OriginCount replaces the page count (1) with a count on estimated groups. It links the read pages by provenance, keeps for each page at most one link to an earlier page, takes the connected components of the kept links as its estimate $\hat{\pi}$ of the origin groups, and reports

$$\hat{N}(a) = \sum_{i=1}^{n} w_i \, s_a(d_i), \qquad w_i = \frac{1}{|\hat{g}(d_i)|} \tag{3}$$

where $\hat{g}(d_i)$ is the estimated group of page d_i. On the true groups $\pi^\star$ the same formula gives the origin count $N_{\mathrm{orig}}(a)$ of Lemma 1. A group contributes the mean support of its pages: one vote when all of its pages state a, and a fraction of a vote when only some of them do. The answer rule, the confidence map f, the threshold τ, and the retrieval policy all stay as they are, and nothing is trained.

Provenance signals. Two read pages are linked when one of three signals applies, in either direction. Level 1 is an explicit reference: one page cites, links, or quotes the other. Level 2 is a shared identifier in the metadata, such as a wire-service tag, a dataset or document identifier, or a canonical URL. Level 3 is restated content: an extractor decides whether one page derives from the other, which a similarity score does not measure. The levels are tried in order, and links are read for the claim that the question asks about, since a page can copy one claim and report another first-hand. Level 3 uses an LLM extractor by default; a rule-based extractor is cheaper and keeps most of the gain (Section 5.4).

Untraced pages and content. A page that no signal links to any other read page forms its own group and keeps a full vote. We call it untraced: it may report a new origin or be a copy with hidden provenance, and the rule treats both cases as new origins instead of dropping the page or attaching it to the nearest group. Every page stays in the context that the agent reads, so a detail that only a copy carries is still kept. Algorithm 1 in Appendix C.4 gives the full procedure and one worked example.

4.2 THEORETICAL ANALYSIS

Two questions remain: which weights keep a count unchanged by copies, and how the count behaves when provenance is incomplete (all proofs are in Appendix B). Consider counts of the form $C = \sum_g \psi(S_g)$, where S_g is the multiset of support values in group g and ψ is the score of a group. The page count, square-root weights, and the count (3) all have this form. A count is *invariant to copies* if adding to a group a page with the same support as all of its pages leaves the count unchanged.

Theorem 1 (One vote per origin)**.** *Let the supports take values in a set $V \subseteq [0,1]$ with at least two elements, and let the score of a group of size m be $\psi(S) = \alpha_m + \gamma_m \sum_{s \in S} s$, which adds up the support of all its pages with equal weight. Then $\psi(\{s\}) = s$ for all $s \in V$ and the count is invariant to copies if and only if $\psi(S) = |S|^{-1} \sum_{s \in S} s$, the mean score of (3). Under this score, adding one more page with support s to a group of m_g pages whose mean support is $\bar{s}_g$ changes the count by $(s - \bar{s}_g)/(m_g + 1)$, while adding it as a group of its own changes the count by s.*

Corollary 1 (Common counts are not invariant to copies)**.** *A copy added to a group of m supporting pages raises the page count by 1, the square-root count $\sum_g |S_g|^{-1/2} \sum_{s \in S_g} s$ by $\sqrt{m+1} - \sqrt{m}$, and a count that weights each page by the reliability $q_i \geq q_{\min} > 0$ of its site by q_i. As copies of one supporting origin accumulate, the confidence under all three tends to 1, while under (3) with the copies traced it stays at $f(1) = \sigma(\beta) \approx 0.587$, the value for one origin.*

Remark. Equal weight inside a group excludes other scores that copies do not change, such as the maximum or the majority of a group, and bounds the influence of a misread page to $1/(m_g + 1)$ of a vote. When all pages of a group share one support, the score is one vote per group, which is also what de-duplication gives a cluster, so OriginCount differs from de-duplication in the signal that forms the groups. Reliability weights can be applied per origin without breaking invariance to copies. Removing copies can take support from a right answer as well as from a wrong one, so accuracy can fall, and one vote per origin leaves correlation between distinct origins in place (see Appendix B.5).

Call a provenance link true when it joins two pages of the same origin. Each page keeps a link to at most one earlier page (Appendix C.4), so $\hat{\pi}$ consists of the connected components of the kept links.

Theorem 2 (Behavior under incomplete provenance)**.** *Let all supports lie in* $[0, 1]$*. (i) If every kept link is true, then* $N_{\text{orig}} \le \hat{N} \le N_{\text{page}}$*, and* $\hat{N} = N_{\text{page}}$ *when no link is kept. (ii) If also every page of group* g *has support* $s_{a,g}$ *and each page but the first read of its group keeps a link with probability* t*, then* $\mathbb{E}[\hat{N}] = N_{\text{page}} - t\,(N_{\text{page}} - N_{\text{orig}})$*. (iii) Merging two groups never raises* $\hat{N}$ *and lowers it by at most one, so if all but* M *kept links are true, then* $N_{\text{orig}} - M \le \hat{N} \le N_{\text{page}}$*.*

Remark. If all links are true, the count can only err towards the page count, and without provenance the agent behaves like the page-count agent. The double counting removed grows linearly with the link rate t of part (ii), and each wrong link costs at most one vote. Keeping an untraced page as its own group preserves part (i); attaching it to the nearest group removes the double counting when the guess is right but can take the count below N_{orig} when the guess is wrong, which Section 5.4 tests.

5 Experiments

5.1 Experimental Setup

Data and agents. We use six fixed web snapshots with page-level provenance labels, two per type of evidence: news (NEWS-A/B), science (SCIENCE-A/B), and product and statistics (PRODUCT-A/B), with 10,000 questions each. All methods get the same pages, so comparisons are paired by question. The default agent is the page-count agent with a Qwen3-32B backbone (Yang et al., 2025) and a dense retriever (Karpukhin et al., 2020). Each question gets eight prefetched retrieval calls, one per round, and twenty pages; the stopping round is the round at which the agent stops (Appendix C.1). About 2.5% of the questions end with an abstention, counted as wrong and not as a confident error.

Baselines and metrics. Table 1 compares 17 baselines and an oracle in five groups of methods: controls, de-duplication by text or by web site (Broder, 1997; Abbas et al., 2023), reliability weighting (Hwang et al., 2025) with a flat confidence discount and a self-reported count (Kadavath et al., 2022), agents with their own stopping rule (Asai et al., 2024; Yan et al., 2024; Jiang et al., 2023; Joren et al., 2025; Li et al., 2025), and evidence aggregation, such as NLI consistency (Honovich et al., 2022), citations (Gao et al., 2023), or conflict weighting (Xu et al., 2024). We report CE, accuracy, the mean stopping round, the expected calibration error (ECE) over two bins split at confidence 0.8 (Naeini et al., 2015; Guo et al., 2017), and the extra cost per read page in milliseconds. Intervals of six-snapshot means are 95% confidence intervals of the mean paired differences (see Appendix C).

5.2 Main Results

In Table 1, OriginCount has the lowest mean CE of all methods except the oracle and the second-lowest ECE (Appendix D.7). Against the page count it removes 5.09 points of CE, with interval $[-5.41, -4.77]$, and lowers ECE from 14.93% to 9.08%. OriginCount has the lowest non-oracle CE on NEWS-B and SCIENCE-A and is within about one point of it on the other four snapshots. It stops 0.29 rounds later than the page count on average, because copies no longer lift its confidence early.

We start with the two common remedies of Section 3. De-duplication and reliability weighting keep CE between 12.7% and 14.0%, and their combination remains 2.41 points above OriginCount (Table 1, bottom). The flat discount has the lowest ECE but leaves CE at 15.88%, since a constant factor lowers the confidence of all answers alike. Random grouping lowers CE by 2.13 points but costs 3.84 points of accuracy, as random merges discard independent evidence. The self-reported count, which asks the reader to count the sources, is the strongest baseline; OriginCount improves on it by 0.97 points of CE and 1.19 points of ECE, and the CE interval of this comparison includes zero.

Accuracy and cost. OriginCount lowers accuracy by 0.99 points; the expected loss still falls once a confident error costs a fifth more than a plain one (Appendix B.6). With the true labels the rule changes mean accuracy by -0.01 points (from -2.42 to $+1.80$ per snapshot), so most of the loss likely comes from grouping errors. With levels 1–2 only, OriginCount has the highest accuracy in the table, 73.46%, and still removes 3.10 points of CE. The extra cost adds 0.32 s per question with the LLM extractor and only 0.04 s with the rule-based one, next to the 2.80 s of retrieval and reading, and the rule-based version still keeps CE at 12.08%, below every baseline in Table 1 (Appendix E.6).

Table 1: Main results at a fixed retrieval budget (Qwen3-32B backbone, dense retriever, eight prefetched calls). Confident-error rate CE (%, ↓) on the six snapshots (N: NEWS, S: SCIENCE, P: PRODUCT), then six-snapshot means of CE, accuracy (Acc, %), stopping round, and expected calibration error ECE (%), and the extra cost per page (ms). L1–2 and L1–3: OriginCount (shaded) with provenance levels 1–2 or 1–3. The oracle (grey) uses the true labels. Best non-oracle CE, accuracy, and ECE in bold, second underlined. †Needs provenance information (links, identifiers, domains, or labels). ‡The discounted confidence never reaches τ. Bottom: paired differences of OriginCount (L1–3) against four methods, with 95% intervals.

	CE per snapshot ↓						Six-snapshot mean				Cost
Method	N-A	N-B	S-A	S-B	P-A	P-B	CE ↓	Acc ↑	Round	ECE ↓	ms
Controls											
Page count (default)	16.72	16.44	18.79	18.71	13.71	14.22	16.43	72.98	3.15	14.93	0.0
Oracle grouping†	11.55	11.50	12.33	12.37	7.02	8.37	10.52	72.97	3.29	9.89	5.0
Random grouping	16.50	14.45	17.31	14.42	9.64	13.45	14.30	69.14	3.34	12.76	3.8
De-duplication											
Verbatim (MinHash)	16.38	14.45	13.90	16.75	7.27	12.08	13.47	72.18	3.40	11.51	0.7
Semantic (embedding)	12.65	12.99	15.65	14.82	11.16	8.98	12.71	71.34	3.37	10.87	4.1
Domain†	12.14	12.39	17.07	17.11	10.66	12.34	13.62	72.19	3.34	11.32	1.0
Reliability and self-report											
Source reliability	14.20	16.07	14.12	14.62	11.98	12.80	13.96	71.94	3.29	10.09	5.7
Flat discount	16.10	15.92	18.05	18.11	13.02	14.08	15.88	72.61	–‡	**8.12**	4.2
Self-reported count	12.14	12.58	16.32	13.38	7.69	11.77	12.31	73.20	3.34	10.27	3.8
Agents with their own stopping rule											
Self-RAG	14.43	14.41	13.27	15.84	13.26	10.38	13.60	72.45	3.42	11.27	5.3
CRAG	14.37	15.98	18.03	17.98	10.74	8.74	14.31	71.59	3.43	13.49	3.8
FLARE	15.22	14.84	15.48	17.87	9.72	10.04	13.86	71.28	3.41	11.81	5.4
Sufficient Context	14.35	15.00	16.22	13.41	7.95	10.84	12.96	73.15	3.35	11.23	4.6
Search-o1	**11.58**	16.14	17.24	16.80	**6.97**	9.56	13.05	72.79	3.31	11.12	5.0
Evidence aggregation											
NLI consistency	12.05	15.59	16.98	15.71	7.12	**8.17**	12.60	73.26	3.38	11.08	5.1
Citations (ALCE)†	14.05	16.15	13.59	13.06	13.25	9.28	13.23	71.44	3.37	11.81	4.8
Conflict weighting	12.38	14.63	18.31	**13.03**	12.89	10.81	13.68	71.38	3.40	11.34	3.6
Reliability + de-dup	13.48	13.49	14.77	14.47	13.13	13.15	13.75	73.25	3.30	11.27	4.9
OriginCount (L1–2)†	12.17	14.40	13.57	17.63	9.08	13.11	13.33	**73.46**	3.33	11.91	7.2
OriginCount (L1–3)†	12.26	**12.18**	**13.23**	13.24	7.94	9.18	**11.34**	71.99	3.44	9.08	16.1

OriginCount (L1–3) minus	ΔCE [95% interval]	ΔAcc [95% interval]	ΔECE [95% interval]
Page count (default)	−5.09 [−5.41, −4.77]	−0.99 [−1.92, −0.06]	−5.85 [−6.74, −4.96]
Self-reported count	−0.97 [−2.52, 0.57]	−1.21 [−3.38, 0.95]	−1.19 [−2.05, −0.33]
Reliability + de-dup	−2.41 [−2.74, −2.09]	−1.26 [−2.09, −0.43]	−2.19 [−2.91, −1.47]
Flat discount	−4.54 [−5.18, −3.90]	−0.62 [−1.08, −0.16]	+0.96 [0.41, 1.51]

Table 2: Generalization and combination. Left and middle: confident-error rate (CE) and accuracy of OriginCount (%) and ΔCE, its change of CE against the page-count agent in the same setting; means over the six snapshots, or over the two snapshots of each type of evidence. Right: CE removed (points) from the page-count agent (16.43%) by each agent variant's own change (Own), by OriginCount used inside the variant without that change (OC), and by both (Both); Sum is Own plus OC before rounding; shaded: default setting.

Backbone	CE	Acc	ΔCE
Qwen3-32B	11.34	71.99	−5.09
Llama-3.3-70B	11.41	71.99	−4.67
Mistral-24B	10.87	71.89	−5.85
gpt-oss-120B	12.17	72.14	−3.77
Qwen3-235B	10.90	71.89	−5.71

Setting	CE	Acc	ΔCE
BM25	11.61	72.03	−4.64
Dense	11.34	71.99	−5.09
Web API	11.39	71.99	−4.69
NEWS	12.22	67.48	−4.36
SCIENCE	13.23	72.38	−5.52
PRODUCT	8.56	76.10	−5.41

Agent variant	Own	OC	Both	Sum
Answer voting	1.82	5.09	6.89	6.91
Page threshold	1.73	5.09	6.73	6.83
FLARE	1.82	4.58	6.43	6.40
Suff. Context	1.63	4.58	6.17	6.21
Logistic gate	1.90	2.82	4.62	4.72

5.3 Generalization and Combination

We vary one factor at a time (Table 2). Across five backbones from 24B to 235B parameters, Qwen3-32B, Llama-3.3-70B (Grattafiori et al., 2024), Mistral-Small-24B (Mistral AI, 2025), gpt-oss-120B

Table 3: Ablations on NEWS-A, SCIENCE-A, and PRODUCT-A. Confident-error rate CE (%, ↓) when one design choice of the full method (levels 1–3, one vote per group, untraced pages kept) is changed, and Δ, the change of the mean. The last row averages over the de-duplication thresholds $\theta \in \{0.7, 0.8, 0.9, 0.95\}$. More results are in Appendix E.

Variant	NEWS-A	SCIENCE-A	PRODUCT-A	Mean ↓	Δ
OriginCount (L1–3)	12.26	13.23	7.94	11.14	
Unit of a group					
Domain instead of origin	12.14	17.07	10.66	13.29	+2.15
Document instead of claim	14.08	15.91	10.31	13.43	+2.29
Embedding clusters instead of provenance	14.42	15.58	10.67	13.56	+2.42
Weight of a group					
Square-root weights instead of one vote	13.31	14.28	9.09	12.23	+1.09
Untraced pages					
Drop them	12.67	13.75	8.49	11.64	+0.50
Attach them to the nearest group	13.47	14.85	9.63	12.65	+1.51
Coverage					
Levels 1–2 only	12.17	13.57	9.08	11.61	+0.47
Group only the top-10 pages	13.18	14.63	9.21	12.34	+1.20
Cheaper variants					
Rule-based extractor instead of LLM	12.44	13.46	8.08	11.33	+0.19
Semantic de-dup first, mean over θ	12.31	13.16	8.02	11.16	+0.02

(OpenAI, 2025), and Qwen3-235B-A22B, OriginCount removes between 3.77 and 5.85 points of CE, and all 30 paired intervals over snapshots and backbones exclude zero (Appendix D.5). The reduction is 4.64 to 5.09 points for the three retrievers and 4.36 to 5.52 points for the three types of evidence. No method, the oracle included, moves by more than 0.75 points of CE between the three retrievers.

We also add OriginCount to five variants of the default agent that change its answer or stopping rule: answer voting, a page threshold variant, FLARE, Sufficient Context, and a logistic stopping gate. They share the backbone, retriever, and budget of the default agent, so their own gains differ from those of the published systems in Table 1. For every variant, the two changes together remove within 0.10 points of the sum of their two separate gains. At twice the retrieval budget, only the oracle has a lower mean CE than OriginCount among the eight methods we ran at that budget (Appendix D.6).

5.4 Ablation Studies

Claim-level origins are a better unit than sites or clusters. In Table 3, grouping by domain costs 2.15 points, grouping whole documents instead of the claim 2.29 points, and clustering by embedding similarity 2.42 points, the most of all variants. Domain grouping merges independent reports from one web site and splits copies across web sites, and embedding clusters face the floor of Proposition 1.

One vote per group gives a lower CE than weights that grow with the group. Square-root weights, which still let a large group outvote a small one, cost 1.09 points of CE, consistent with Corollary 1.

Keeping untraced pages works better than guessing their origin. Dropping untraced pages costs 0.50 points of CE and attaching them to the nearest group 1.51 points, about three times as much on these three and more than twice as much on each of the six snapshots (Appendix E.5). A wrongly attached page is a false link, which can take the count below the true origin count (Theorem 2(iii)).

Level 3 and full coverage help, and a cheap extractor keeps most of the gain. Dropping level 3 costs 0.47 points here and 1.99 over all six snapshots (Table 1), and grouping only the top ten pages costs 1.20 points. A rule-based extractor costs 0.19 points here and 0.74 over all six at one eighth of the extra cost, and a semantic de-duplication step before grouping changes CE by only 0.02 points.

Provenance is less affected by noise than similarity. Under noise in the grouping signals on pools with seven origins and known provenance, the provenance rule stays within one origin of the true count at every level, while clustering by embedding similarity overcounts by 6.3 origins on average at the highest level, and a hybrid of the two lies in between once noise is added (see Appendix E.7).

Copies with new facts raise CE a little. When half of all the copies carry a new fact and are merged anyway, which breaks part (i) of our Assumption 1, CE rises by at most 0.94 points, and discarding the clarifying content that the grouped pages add raises it by 0.48 points on average (Appendix E.9).

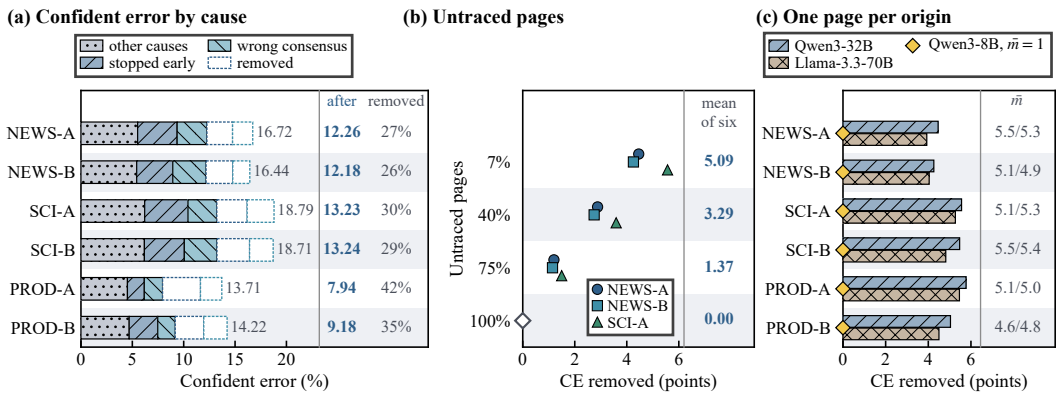


Figure 2: Where the reduction comes from. (a) Confident errors of the page-count agent by cause, with its CE at the end of each bar. Filled segments remain under OriginCount; dashed white segments are removed (blue: stopped early, teal: wrong consensus). Other causes are retrieval misses (no read page states the correct answer) and errors of extraction, aggregation, or other kinds. Right: CE of OriginCount and the share removed. (b) CE removed on three snapshots and its mean over all six (right) as provenance is hidden from more pages; the hollow diamond marks the case without provenance. (c) CE removed with Qwen3-32B and Llama-3.3-70B, and a Qwen3-8B control in which every origin gives one page (diamonds). Right: pages per origin $\bar{m}$ over all read pages (Qwen3-32B/Llama-3.3-70B). SCI and PROD abbreviate SCIENCE and PRODUCT.

5.5 Error Analysis

OriginCount changes only the count, which enters the agent in two places, the stopping rule and the vote for an answer. Fig. 2a splits the confident errors of the page-count agent by cause. The errors that it removes come from stopping too early and from a wrong consensus, in which copies make a wrong answer look well supported. Errors from other causes, such as a correct answer that no read page states, need a fix elsewhere in the agent. On each snapshot, OriginCount removes 39% to 69% of the confident errors from early stopping and 34% to 57% of those from a wrong consensus, or 26% to 42% of all confident errors, and the other causes make up about half of the errors that remain.

Fig. 2b hides provenance from a growing share of the retrieved pages. As the untraced share rises from 7% to 40%, 75%, and 100%, the reduction shrinks from 5.09 to 3.29, 1.37, and 0 points as fewer links are kept (Appendix E.5), and without provenance OriginCount returns the output of the page-count agent. On noise-free controlled pools, the provenance rule merges no pages of different origins and so does not undercount, in line with Theorem 2(i), while a citation-only rule makes a few such merges and undercounts by one or two origins when all links are observed (Appendix E.8). In Fig. 2c, OriginCount removes 3.93 to 5.77 points per snapshot with Qwen3-32B and Llama-3.3-70B. In a control where every origin contributes one page, there are no copies to group and thus no reduction, while under standard retrieval the same Qwen3-8B backbone gains 3.7 to 6.1 points of CE.

6 Conclusion

Many retrieval agents count retrieved pages, and on six web snapshots most retrieved pages repeat an origin that another read page reports. A copy that depends on the answer only through its origin adds no new evidence about the answer, yet reliability weighting counts it, and de-duplication often does too. Among counts that add up page support with equal weight inside a group and count a lone page as one vote, one vote per provenance group is the only one that copies do not change. OriginCount applies this count with three provenance signals and keeps each untraced page as its own group. It needs no training and replaces only the counting step, so any agent that counts support can use it, and a rule-based extractor keeps most of its benefit at 2 ms per page. With a Qwen3-32B agent at the fixed budget, it removes 31% of the confidently wrong answers made by the page-count agent at a cost of about one point of accuracy. It lowers CE on five backbones, three retrievers, three types of evidence, and five agent variants, and falls back to the page count when no provenance can be traced.

Limitations. The gain depends on the provenance signals that pages expose and shrinks as fewer pages can be traced. One vote per origin also does not remove correlation between distinct origins, a different kind of dependence, such as two laboratories that use one flawed instrument (Appendix B.5).

Reproducibility Statement

The counting rule is specified by (3) and Algorithm 1, and Appendix C.4 describes the three provenance levels, the extractors, and the policy for untraced pages. Appendix B contains all proofs. Appendix C lists the snapshots, the agents, the backbones, the retrievers, the baselines, the metrics, and the intervals, and Appendices D and E report the numbers behind the main text. We will release the snapshots with their labels, the evaluation code, and the scripts that produce the tables and figures.

AI Use Statement

Generative AI tools were used only to edit the wording of text written by the authors, to search for related work and check bibliographic details, and to write code for plots, tables and typesetting. They were not used for any task that the ICLR 2027 AI policy requires authors to disclose: developing theory, formulating or proving mathematical claims, proposing or refining hypotheses, designing, implementing or running the method or experiments, generating, cleaning or analyzing data, translating text, or interpreting results. The models in our experiments are objects of study, not writing aids. The authors reviewed all AI-assisted work and take full responsibility for the content of this paper.

References

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

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.

Sven Berg. Condorcet's jury theorem, dependency among jurors. *Social Choice and Welfare*, 10(1): 87–95, 1993.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022.

Philip J. Boland. Majority systems and the Condorcet jury theorem. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 38(3):181–189, 1989.

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

Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. Syntactic clustering of the Web. *Computer Networks and ISDN Systems*, 29(8–13):1157–1166, 1997.

Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and where: A characterization of data provenance. In *Database Theory – ICDT 2001*, volume 1973 of *Lecture Notes in Computer Science*, pp. 316–330. Springer, 2001.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998.

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

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. The power of noise: Redefining retrieval for RAG systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 719–729, 2024.

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.

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.

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):1358–1369, 2010.

Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proceedings of the VLDB Endowment*, 8(9):938–949, 2015.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3905–3920, 2022.

Stuart H. Hurlbert. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, 54(2):187–211, 1984.

Jeongyeon Hwang, Junyoung Park, Hyejin Park, Dongwoo Kim, Sangdon Park, and Jungseul Ok. Retrieval-augmented generation with estimation of source reliability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34279–34303, 2025.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, 2024.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, 2023.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O. Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Conference on Language Modeling*, 2025.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, 2025.

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, 1997.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696, 2020.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781, 2020.

Leslie Kish. *Survey Sampling*. John Wiley & Sons, New York, 1965.

Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

Dahyun Lee, Yongrae Jo, Haeju Park, and Moontae Lee. Shifting from ranking to set selection for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17606–17619, 2025.

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.

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 497–506, 2009.

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*, volume 33, pp. 9459–9474, 2020.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5420–5438, 2025.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. A survey on truth discovery. *ACM SIGKDD Explorations Newsletter*, 17(2):1–16, 2016.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, 2023.

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*, pp. 141–150, 2007.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

Mistral AI. Mistral small 3.1. https://mistral.ai/news/mistral-small-3-1, 2025. Blog post, March 17, 2025.

Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

Ravali Pochampally, Anish Das Sarma, Xin Luna Dong, Alexandra Meliou, and Divesh Srivastava. Fusing data with correlations. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pp. 433–444, 2014.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12991–13013, 2024.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

Martin R. Tramèr, D. John M. Reynolds, R. Andrew Moore, and Henry J. McQuay. Impact of covert duplicate publication on meta-analysis: A case study. *BMJ*, 315(7109):635–640, 1997.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, 2023.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, 2009.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

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*, 2024.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565, 2024.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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.

Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Truth discovery with multiple conflicting information providers on the Web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808, 2008.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4615–4635, 2023.

Appendix

A Notation 16

B Proofs and Derivations 16

B.1 Proof of Lemma 1 . . . 17
B.2 Proof of Proposition 1 . . . 18
B.3 Proof of Theorem 1 and Corollary 1 . . . 18
B.4 Proof of Theorem 2 . . . 19
B.5 Correlated Origins . . . 20
B.6 Effect of the Count on Accuracy . . . 20

C Experimental Details 21

C.1 Snapshots and Protocol . . . 21
C.2 Agents, Backbones, and Retrievers . . . 21
C.3 Baselines . . . 22
C.4 Implementation of OriginCount . . . 22
C.5 Metrics and Intervals . . . 23
C.6 Controlled Experiments . . . 23

D Additional Results 24

D.1 Accuracy and Stopping Round per Snapshot . . . 24
D.2 Intervals on NEWS-A . . . 24
D.3 All Methods under Three Retrievers . . . 24
D.4 All Methods under Five Backbones . . . 24
D.5 Paired Differences per Snapshot and Backbone . . . 25
D.6 Retrieval Budget . . . 25
D.7 Calibration . . . 25
D.8 Extra Cost and Retrievers . . . 25
D.9 Accuracy and Confident Errors . . . 26

E Additional Ablation Studies 27

E.1 Grouping Quality against the Provenance Labels . . . 27
E.2 Controlled Copies . . . 28
E.3 Decomposition of the CE Reduction . . . 28
E.4 De-duplication . . . 29
E.5 Untraced Pages . . . 29
E.6 Provenance Levels, Extractors, and Cost . . . 30
E.7 Noise in the Grouping Signals . . . 30
E.8 Partially Observed Provenance . . . 30
E.9 Copies with New Facts and Discarded Content . . . 30
E.10 Combination with Other Agents . . . 31
E.11 Experiments with Correlated Origins . . . 31

F Comparison of Counting Units **31**

G Values Plotted in the Figures **36**

A Notation

Table 4 lists the symbols of the paper and where each is first defined. Counts are written N with a subscript for their unit, such as pages or origins, and, apart from the answer $\hat{a}$, a hat marks a quantity, such as $\hat{N}$, $\hat{G}$, or $\hat{\pi}$, that OriginCount estimates from the pages instead of reading it from the labels.

Table 4: Symbols used in the paper and where they are defined.

Symbol	Meaning	Defined
x, a, $\hat{a}$, H_a	question, candidate answer, answer of the agent, event that a is correct	§3.1
d_i, n	a read page and the number of pages read	§3.1
$s_a(d_i)$	support of d_i for a: 1 if the page states a and 0 otherwise (the theorems allow $[0, 1]$)	§3.1
$N(a)$, $N_{\text{page}}(a)$	support count of the agent, and the page count (1)	§3.1
f, σ, β, c	confidence map $f(N) = \sigma(\beta N)$, logistic function, slope 0.35, confidence	§3.1
τ, N_τ	stopping threshold 0.9, and the count $f^{-1}(\tau)$ at which it is reached	§3.1, Lem. 1
$o(d)$, g, $\pi^\star$	origin of page d, group of pages with one origin, partition of the pages into groups	§3.1
m_g, G, $\bar{m}$	size of group g, number of groups, pages per origin	§3.1, Fig. 2
k	number of pages added to one origin in the controlled pools	Fig. 1
E_g, D_g	origin record and copies of group g	Ass. 1
L_0, ℓ_g, Γ	prior log-odds given $\pi^\star$, log-likelihood ratio of E_g, gap (2)	Lem. 1
$s_{a,g}$, $N_{\text{orig}}(a)$	common support of the pages of group g, and the origin count	Lem. 1, §4.1
$\hat{\pi}$, $\hat{g}(d_i)$, w_i	estimated groups, estimated group of d_i, weight $1/\lvert\hat{g}(d_i)\rvert$	§4.1
$\hat{N}(a)$, $\hat{G}$	count (3) on the estimated groups, and the number of estimated groups	§4.1, App. E.1
S_g, $\bar{s}_g$, ψ	multiset of supports in group g, their mean, and the score of a group	§4.2
V, α_m, γ_m	set of support values, and the intercept and slope of ψ for groups of size m	Thm. 1
q_i, $q_{\min}$	reliability weight of the site of d_i, and a lower bound on these weights	Cor. 1
Z, F_{copy}, F_{ind}	score of a page pair, its distribution on copy pairs and on independent pairs	Prop. 1
FS, FM, TV	false-split rate, false-merge rate, total variation distance	Prop. 1
t, M, k_g	chance that a page after the first of its group keeps a link, number of false kept links, number of groups of $\hat{\pi}$ inside group g	Thm. 2, App. B.4
ρ, G_{eff}	correlation between distinct origins, effective number of origins	App. B.5
R, B, λ	shares of questions moved from wrong to right and from right to wrong, extra loss of a confident error	App. B.6
θ	similarity threshold of semantic de-duplication	Table 3
CE, ECE	confident-error rate and expected calibration error, in %	§3.1, §5.1

B Proofs and Derivations

Throughout, $\neg H_a$ is the complement of H_a, $\sigma(z) = 1/(1+e^{-z})$ is the standard logistic function, and $E_{1:G}$ and $D_{1:G}$ collect the origin records and the copies of all G groups. Every statement conditions on the partition $\pi^\star$, as Assumption 1 does, and the supports of a fixed candidate a are written simply $s(d)$. We also assume $0 < \mathbb{P}(H_a \mid E_g) < 1$ for every group, so that every ratio below is well defined.

B.1 Proof of Lemma 1

Step 1: the copies drop out. Part (i) of Assumption 1 also holds under $\neg H_a$. Indeed $\mathbb{P}(D_g \mid E_g)$ is a mixture of $\mathbb{P}(D_g \mid E_g, H_a)$ and $\mathbb{P}(D_g \mid E_g, \neg H_a)$ with positive weights, and when one component equals the mixture, so does the other one. With Bayes' rule and part (ii), the posterior odds of H_a factor over the groups:

$$\begin{aligned}\frac{\mathbb{P}(H_a \mid E_{1:G}, D_{1:G})}{\mathbb{P}(\neg H_a \mid E_{1:G}, D_{1:G})} &= \frac{\mathbb{P}(H_a)}{\mathbb{P}(\neg H_a)} \prod_{g=1}^{G} \frac{\mathbb{P}(E_g, D_g \mid H_a)}{\mathbb{P}(E_g, D_g \mid \neg H_a)} \\ &= \frac{\mathbb{P}(H_a)}{\mathbb{P}(\neg H_a)} \prod_{g=1}^{G} \frac{\mathbb{P}(E_g \mid H_a)\,\mathbb{P}(D_g \mid E_g, H_a)}{\mathbb{P}(E_g \mid \neg H_a)\,\mathbb{P}(D_g \mid E_g, \neg H_a)} \\ &= \frac{\mathbb{P}(H_a)}{\mathbb{P}(\neg H_a)} \prod_{g=1}^{G} \frac{\mathbb{P}(E_g \mid H_a)\,\mathbb{P}(D_g \mid E_g)}{\mathbb{P}(E_g \mid \neg H_a)\,\mathbb{P}(D_g \mid E_g)} \\ &= \frac{\mathbb{P}(H_a)}{\mathbb{P}(\neg H_a)} \prod_{g=1}^{G} \frac{\mathbb{P}(E_g \mid H_a)}{\mathbb{P}(E_g \mid \neg H_a)} = \frac{\mathbb{P}(H_a)}{\mathbb{P}(\neg H_a)} \prod_{g=1}^{G} e^{\ell_g}. \end{aligned} \tag{4}$$

The factor $\mathbb{P}(D_g \mid E_g)$ appears in the numerator and the denominator and cancels, so the copies leave the posterior unchanged. Taking logarithms gives the posterior log-odds $L_0 + \sum_g \ell_g$ of the lemma.

Step 2: the gap. A count that credits each of the m_g pages of group g with ℓ_g has log-odds $L_0 + \sum_g m_g \ell_g$. When the pages of each group share one support $s_{a,g}$, we have $N_{\text{page}}(a) = \sum_g m_g s_{a,g}$ and $N_{\text{orig}}(a) = \sum_g s_{a,g}$, and with $\ell_g = \beta s_{a,g}$ the difference of the two log-odds is

$$\begin{aligned}\Gamma &= \Big(L_0 + \sum_{g=1}^{G} m_g \ell_g\Big) - \Big(L_0 + \sum_{g=1}^{G} \ell_g\Big) = \sum_{g=1}^{G} (m_g - 1)\,\ell_g \\ &= \beta \sum_{g=1}^{G} m_g\, s_{a,g} - \beta \sum_{g=1}^{G} s_{a,g} \\ &= \beta \big(N_{\text{page}}(a) - N_{\text{orig}}(a)\big). \end{aligned} \tag{5}$$

Each term of Γ has the sign of ℓ_g and the factor $m_g - 1 \geq 0$. When every origin supports a, the gap is therefore positive once a group holds a copy, and it grows with the evidence ℓ_g of the copied origin.

Step 3: the stopping round. Let supporting origins arrive one per round, each with m pages of support one. After r rounds, $N_{\text{orig}} = r$ and $N_{\text{page}} = mr$. The agent stops at the first round with $f(N) \geq \tau$, that is with $N \geq N_\tau = f^{-1}(\tau) = \beta^{-1} \log \frac{\tau}{1-\tau}$, so

$$r_{\text{page}} = \min\{r : mr \geq N_\tau\} = \Big\lceil \frac{N_\tau}{m} \Big\rceil, \qquad r_{\text{orig}} = \min\{r : r \geq N_\tau\} = \lceil N_\tau \rceil. \tag{6}$$

Under the page count the agent thus reads about $1/m$ as many origins before it stops. □

When the origin record is not read. Suppose that the origin record of group g is not read and that all read pages of g depend on H_a only through E_g; we write D_g for them and write $\ell(e) = \log \frac{\mathbb{P}(e \mid H_a)}{\mathbb{P}(e \mid \neg H_a)}$ for a possible origin record e. We sum over the records and use part (i) and Step 1 to replace $\mathbb{P}(D_g \mid e, H_a)$ and $\mathbb{P}(D_g \mid e, \neg H_a)$ by $\mathbb{P}(D_g \mid e)$:

$$\begin{aligned}\frac{\mathbb{P}(D_g \mid H_a)}{\mathbb{P}(D_g \mid \neg H_a)} &= \frac{\sum_e \mathbb{P}(D_g \mid e)\,\mathbb{P}(e \mid H_a)}{\sum_{e'} \mathbb{P}(D_g \mid e')\,\mathbb{P}(e' \mid \neg H_a)} = \frac{\sum_e \mathbb{P}(D_g \mid e)\,\mathbb{P}(e \mid \neg H_a)\, e^{\ell(e)}}{\sum_{e'} \mathbb{P}(D_g \mid e')\,\mathbb{P}(e' \mid \neg H_a)} \\ &= \sum_e \mathbb{P}(e \mid D_g, \neg H_a)\, e^{\ell(e)} \\ &= \mathbb{E}\big[e^{\ell(E_g)} \,\big|\, D_g, \neg H_a\big] \;\leq\; \max_e e^{\ell(e)}, \end{aligned} \tag{7}$$

where the maximum runs over the records with $\mathbb{P}(e \mid D_g, \neg H_a) > 0$. The copies of the group together move the log-odds by at most $\max_e \ell(e)$, which is the evidence of a single record, however many copies of the origin were read. When $m_g \geq 2$ and a record carries at most the unit β of evidence, a count that gives each of these pages its own unit of evidence still overstates this group.

B.2 Proof of Proposition 1

Let a rule merge a pair with score $Z = z$ with probability $\varphi(z) \in [0, 1]$; a threshold θ gives $\varphi(z) = \mathbb{1}\{z \geq \theta\}$. The false-split rate is the chance of keeping a copy pair apart, $\mathrm{FS} = \mathbb{E}_{F_{\mathrm{copy}}}[1-\varphi(Z)]$, and the false-merge rate is the chance of joining an independent pair, $\mathrm{FM} = \mathbb{E}_{F_{\mathrm{ind}}}[\varphi(Z)]$. Let p_{c} and p_{i} be the densities of F_{copy} and F_{ind} with respect to $\nu = F_{\mathrm{copy}} + F_{\mathrm{ind}}$, and let $A^\star = \{z : p_{\mathrm{c}}(z) > p_{\mathrm{i}}(z)\}$. Then

$$\begin{aligned}
\mathrm{FS} + \mathrm{FM} &= \mathbb{E}_{F_{\mathrm{copy}}}\big[1 - \varphi(Z)\big] + \mathbb{E}_{F_{\mathrm{ind}}}\big[\varphi(Z)\big] \\
&= 1 - \int \varphi(z)\,\big(p_{\mathrm{c}}(z) - p_{\mathrm{i}}(z)\big)\,d\nu(z) \\
&\geq 1 - \int_{A^\star} \big(p_{\mathrm{c}}(z) - p_{\mathrm{i}}(z)\big)\,d\nu(z) \\
&= 1 - \sup_{A} \big|F_{\mathrm{copy}}(A) - F_{\mathrm{ind}}(A)\big| \\
&= 1 - \mathrm{TV}(F_{\mathrm{copy}}, F_{\mathrm{ind}}).
\end{aligned} \tag{8}$$

The inequality holds because $0 \leq \varphi \leq 1$, so the integral is largest when φ is one on $A^\star$ and zero elsewhere. The last two lines are the definition of the total variation distance. The rule $\varphi = \mathbb{1}_{A^\star}$ attains the bound, so the bound cannot be improved (Le Cam, 1973; Tsybakov, 2009). Finally $\max(\mathrm{FS}, \mathrm{FM}) \geq (\mathrm{FS} + \mathrm{FM})/2 \geq (1 - \mathrm{TV})/2$. □

Consequence for de-duplication. If all pages of a group share one support, keeping one page per cluster gives each cluster one vote, which is (3) on the clusters, so Theorem 2 applies. If each page after the first of its group keeps a link with probability t and no false merge occurs, part (ii) gives the expected count $N_{\mathrm{page}} - t\,(N_{\mathrm{page}} - N_{\mathrm{orig}})$, so the copies left in place still add $(1-t)(N_{\mathrm{page}} - N_{\mathrm{orig}})$, and a threshold that splits more of the copy pairs has a smaller t. Each false merge can remove up to one vote by part (iii). Proposition 1 says that the two errors cannot both be small once rewriting brings the scores of copy pairs close to those of independent pairs, which is consistent with Fig. 6b.

The distance for provenance signals. The bound also holds for the same-group decision of OriginCount, taken as the score Z of a pair. With all three levels, Table 18 gives $\mathrm{FS} = 14.6\%$ on same-origin and $\mathrm{FM} = 1.24\%$ on different-origin pairs, so the argument of (8), applied to these two kinds of page pair, gives a total variation distance of at least 0.84 between the two for this decision.

B.3 Proof of Theorem 1 and Corollary 1

Let $s_1 \neq s_2$ be two values in V, which exist because V has at least two elements. We first show that the two conditions of the theorem force the score to be the mean, and then that the mean meets both.

Only the mean meets both conditions. A single page gives $\alpha_1 + \gamma_1 s = s$ for $s \in \{s_1, s_2\}$, so $\gamma_1 = 1$ and $\alpha_1 = 0$. Invariance to copies says that adding a page of support s to a group of m pages of support s leaves the score unchanged, for every $m \geq 1$ and every $s \in V$. We write this condition at $s = s_1$ and at $s = s_2$, subtract the two equations, insert the result back into either one, and unroll the recursion:

$$\begin{aligned}
&\alpha_{m+1} + (m+1)\,\gamma_{m+1}\,s = \alpha_m + m\,\gamma_m\,s \qquad \text{for } s \in \{s_1, s_2\} \\
\Longrightarrow\quad &\big((m+1)\,\gamma_{m+1} - m\,\gamma_m\big)(s_1 - s_2) = 0 \quad \text{and} \quad \alpha_{m+1} = \alpha_m \\
\Longrightarrow\quad &m\,\gamma_m = (m-1)\,\gamma_{m-1} = \cdots = \gamma_1 = 1, \qquad \alpha_m = \alpha_{m-1} = \cdots = \alpha_1 = 0 \\
\Longrightarrow\quad &\psi(S) = \alpha_{|S|} + \gamma_{|S|} \sum_{s \in S} s = \frac{1}{|S|} \sum_{s \in S} s.
\end{aligned} \tag{9}$$

Since $s_1 - s_2 \neq 0$, the second line gives $(m+1)\gamma_{m+1} = m\gamma_m$, and then either of the two equations in the first line gives $\alpha_{m+1} = \alpha_m$. Both coefficients are therefore fixed for every group size $m \geq 1$.

The mean meets both conditions. For $\psi(S) = |S|^{-1} \sum_{s \in S} s$, a group with a single page scores s, and adding a page of support s to m pages of support s gives $(ms + s)/(m+1) = s$, which equals the old score. Summing the score over the groups gives the count (3), which proves the first claim.

Influence of one page. Let group g hold m_g pages with mean support $\bar{s}_g$, and add a page with support s. The score of the group changes by

$$\frac{m_g \bar{s}_g + s}{m_g + 1} - \bar{s}_g = \frac{m_g \bar{s}_g + s - (m_g + 1)\,\bar{s}_g}{m_g + 1} = \frac{s - \bar{s}_g}{m_g + 1}, \tag{10}$$

which lies in $[-\frac{1}{m_g+1}, \frac{1}{m_g+1}]$ because s and $\bar{s}_g$ lie in $[0, 1]$. A page that starts a new group adds $\psi(\{s\}) = s$. □

Proof of Corollary 1. A group of m supporting pages scores m under the page count, $m^{-1/2} \cdot m = \sqrt{m}$ under the square-root count, and $\sum_i q_i$ under reliability weights. One more supporting copy therefore adds 1, $\sqrt{m+1} - \sqrt{m} > 0$, and $q_i \geq q_{\min}$ to the three counts. With m pages of one supporting origin, the counts are m, $\sqrt{m}$, and at least $q_{\min} m$, which all grow without bound, so $f(N) = \sigma(\beta N)$ tends to one. Under (3) with the copies traced, the count stays at one and

$$f(1) = \sigma(0.35) = \frac{1}{1 + e^{-0.35}} = \frac{1}{1 + 0.7047} = 0.5866. \tag{11}$$

Table 30 in Appendix F gives the limit for the main counting units of the paper. □

B.4 Proof of Theorem 2

The kept links split the read pages into connected components, the groups of $\hat{\pi}$. For a component P we write $\bar{p}$ for the mean support of its pages, so that $\hat{N} = \sum_P \bar{p}$ with the sum over all components.

Part (i). If every kept link is true, every component lies inside one true group. Let the components inside group g be $P_{g,1}, \ldots, P_{g,k_g}$ with means $\bar{p}_{g,1}, \ldots, \bar{p}_{g,k_g}$. The mean support of the whole group is a weighted mean of these means, and supports are non-negative, so

$$\begin{aligned} \frac{1}{m_g} \sum_{d \in g} s(d) = \sum_{h=1}^{k_g} \frac{|P_{g,h}|}{m_g}\, \bar{p}_{g,h} \;&\leq\; \max_h \bar{p}_{g,h} \;\leq\; \sum_{h=1}^{k_g} \bar{p}_{g,h} \\ &\leq \sum_{h=1}^{k_g} |P_{g,h}|\, \bar{p}_{g,h} = \sum_{d \in g} s(d). \end{aligned} \tag{12}$$

Summing over the groups, the left side gives N_{orig}, the middle term then gives $\hat{N}$, and the right side N_{page}, so $N_{\text{orig}} \leq \hat{N} \leq N_{\text{page}}$. With no kept link every page is its own component and $\hat{N} = N_{\text{page}}$.

Part (ii). Now every page of group g has support $s_{a,g}$, so every component inside g has mean $s_{a,g}$ and $\hat{N} = \sum_g k_g s_{a,g}$. Each page keeps at most one link, to an earlier page, so the kept links inside g form a forest, and a forest has as many components as nodes minus edges. Hence $k_g = m_g - X_g$, where X_g is the number of pages of g that keep a link. The first page read of g keeps none, and each of the other $m_g - 1$ pages keeps one with probability t, so $\mathbb{E}[X_g] = t\,(m_g - 1)$ by linearity, whether or not the pages are linked independently. Hence

$$\begin{aligned} \mathbb{E}[k_g] &= m_g - \mathbb{E}[X_g] = m_g - t\,(m_g - 1), \\ \mathbb{E}[\hat{N}] &= \sum_{g=1}^{G} s_{a,g}\, \mathbb{E}[k_g] = \sum_{g=1}^{G} s_{a,g} \big(m_g - t\,(m_g - 1)\big) \\ &= \sum_{g=1}^{G} m_g\, s_{a,g} - t \sum_{g=1}^{G} (m_g - 1)\, s_{a,g} \\ &= N_{\text{page}} - t\,(N_{\text{page}} - N_{\text{orig}}). \end{aligned} \tag{13}$$

When the pages are linked independently of each other, the variance of $\hat{N}$ is $t\,(1-t) \sum_g (m_g - 1)\, s_{a,g}^2$.

Part (iii). Let two groups P and P' with means $\bar{p}$ and $\bar{p}'$ be merged. The merged group has mean $(|P|\bar{p} + |P'|\bar{p}')/(|P| + |P'|)$, so the count changes by

$$\begin{aligned}\Delta &= \frac{|P|\,\bar{p} + |P'|\,\bar{p}'}{|P| + |P'|} - \big(\bar{p} + \bar{p}'\big) \\ &= -\frac{|P'|\,\bar{p} + |P|\,\bar{p}'}{|P| + |P'|} \;\in\; \big[-\max(\bar{p}, \bar{p}'),\, 0\big] \subseteq [-1, 0], \end{aligned} \tag{14}$$

since the fraction is a weighted mean of $\bar{p}$ and $\bar{p}'$. Merging therefore never raises $\hat{N}$ and lowers it by at most one. Now let all kept links but M be true. Without these M links, the kept links split the pages into finer groups that each lie inside one true group, and part (i) gives $N_{\text{orig}} \le \hat{N} \le N_{\text{page}}$ for them. Adding the M links back merges groups at most M times, which lowers the count by at most M and never raises it, so $N_{\text{orig}} - M \le \hat{N} \le N_{\text{page}}$. □

B.5 Correlated Origins

Assumption 1(ii) asks distinct origins to be independent. When they are not, one vote per origin still overstates the evidence, and no weight inside a group can change this. Let $Y_1, \ldots, Y_G \in \{0, 1\}$ indicate whether each of G distinct origins is correct, with common variance v and pairwise correlation ρ. The mean vote $\bar{Y}$ has variance

$$\begin{aligned}\mathrm{Var}(\bar{Y}) &= \frac{1}{G^2}\Big(\sum_{g=1}^{G} \mathrm{Var}(Y_g) + \sum_{g \neq g'} \mathrm{Cov}(Y_g, Y_{g'})\Big) \\ &= \frac{1}{G^2}\big(G\,v + G(G-1)\,\rho\,v\big) \\ &= \frac{v}{G}\big(1 + (G-1)\,\rho\big) = \frac{v}{G_{\text{eff}}}, \qquad G_{\text{eff}} = \frac{G}{1 + (G-1)\,\rho}, \end{aligned} \tag{15}$$

which is the design effect of cluster sampling (Kish, 1965). The vote of G correlated origins is thus as variable as that of G_{eff} independent ones. The weights of (3) act inside groups and leave ρ unchanged. Table 5 then evaluates (15) for $G = 5$, and Appendix E.11 measures this effect on controlled pools.

Table 5: Effective number of independent origins from (15) at $G = 5$, and the share of the gain of pooling five origins that the correlation takes back, $(5 - G_{\text{eff}})/(5 - 1)$.

ρ	0	0.2	0.5	0.8
$1 + (G-1)\rho$	1.0	1.8	3.0	4.2
G_{eff}	5.00	2.78	1.67	1.19
Share taken back	0%	56%	83%	95%

B.6 Effect of the Count on Accuracy

Let $\hat{a}_{\text{p}}$ and $\hat{a}_{\text{o}}$ be the answers of the agent under the page count and under OriginCount, and let $a^\star$ be the correct answer. Splitting each accuracy by the outcome under the other count gives

$$\begin{aligned}\mathrm{Acc}_{\text{o}} - \mathrm{Acc}_{\text{p}} &= \mathbb{P}(\hat{a}_{\text{o}} = a^\star) - \mathbb{P}(\hat{a}_{\text{p}} = a^\star) \\ &= \mathbb{P}(\hat{a}_{\text{o}} = a^\star, \hat{a}_{\text{p}} \neq a^\star) + \mathbb{P}(\hat{a}_{\text{o}} = a^\star, \hat{a}_{\text{p}} = a^\star) \\ &\quad - \mathbb{P}(\hat{a}_{\text{p}} = a^\star, \hat{a}_{\text{o}} \neq a^\star) - \mathbb{P}(\hat{a}_{\text{p}} = a^\star, \hat{a}_{\text{o}} = a^\star) \\ &= R - B, \end{aligned} \tag{16}$$

where R and B are the shares of questions moved from wrong to right and from right to wrong. On the same read pages, the leader a_1 and the runner-up a_2 of the page count swap only if the count removed from a_1 exceeds that removed from a_2 by more than the lead, $\Delta_{a_1} - \Delta_{a_2} > N_{\text{page}}(a_1) - N_{\text{page}}(a_2)$, with $\Delta_a = N_{\text{page}}(a) - \hat{N}(a)$. Assumption 1 says how copies arise but not which candidate they support, so the sign of $R - B$ depends on the data. Appendix D.9 reports R and B for each snapshot.

The cost of a confident error. Suppose a deployment loses one unit for every wrong answer and $\lambda \ge 0$ more units when the wrong answer is also confident. The expected loss in points is

$\mathcal{L} = \lambda\,\mathrm{CE} + (100 - \mathrm{Acc})$, and moving from the page count to OriginCount changes it by

$$\Delta\mathcal{L} = \lambda\,(11.34 - 16.43) + (72.98 - 71.99) = -5.09\,\lambda + 0.99, \tag{17}$$

which is negative for every $\lambda > 0.195$. So OriginCount has the lower expected loss once a confidently wrong answer costs the deployment about a fifth more than a wrong answer given with low confidence.

C Experimental Details

C.1 Snapshots and Protocol

The six snapshots are fixed web corpora, two for each type of evidence: news events (NEWS-A and NEWS-B), scientific data (SCIENCE-A and SCIENCE-B), and product and statistics questions (PRODUCT-A and PRODUCT-B). Each has 10,000 questions, and every retrieved page carries a provenance label that names its origin. The labels define the true groups $\pi^\star$ that the oracle uses and against which Appendix E.1 scores the estimated groups of pages. Because the corpora are fixed, every method gets the same pages for the same question, and all comparisons are paired by question.

Every question gets eight prefetched retrieval calls, one per round, with twenty pages in all, and a context budget of 4,096 tokens. The stopping round is the first round at which the confidence reaches τ, or at which an agent with its own stopping rule stops, and its mean is taken over the questions that stop; the other questions are answered after the last call. Since the calls are made in advance for all methods alike, a later stop adds no call and an earlier stop saves none, so the stopping round measures how much evidence the agent reads before it answers. About 2.5% of the questions end in an abstention, which counts as wrong, is never a confident error, and falls in neither bin of the ECE.

C.2 Agents, Backbones, and Retrievers

The default agent is the page-count agent of Section 3.1: it reads the pages round by round, answers the candidate with the largest count, reports $\sigma(0.35N)$, and stops at $\tau = 0.9$. Section 5.3 adds OriginCount to five variants of this agent. Two count support and act on the count directly, a page threshold variant and answer voting (Wang et al., 2023). The other three decide when to stop in their own way: Sufficient Context asks whether the context is enough (Joren et al., 2025), FLARE retrieves when the confidence of the generation drops (Jiang et al., 2023), and a logistic gate makes the stopping decision. Each of them uses a support count at one point of its decision, and this is where OriginCount replaces the page count. Table 6 lists these agents together with Self-RAG and CRAG, and it checks that without any provenance OriginCount leaves the output of every agent unchanged.

Table 6: Agent check. Type (counting or deciding), share of untraced pages, and abstention rate for each agent, and the largest change of CE, accuracy, or ECE when OriginCount runs without any provenance, against the page count of the same agent.

Agent	Type	Untraced (%)	Abstain (%)	Largest change
Page threshold	counting	6.7	2.5	0
Answer voting	counting	7.3	2.7	0
Sufficient Context	deciding	6.5	2.4	0
Logistic gate	deciding	7.8	2.9	0
FLARE	deciding	6.9	2.6	0
Self-RAG	deciding	7.1	2.2	0
CRAG	deciding	7.6	2.8	0

The default reader is Qwen3-32B (Yang et al., 2025). Section 5.3 replaces it with Llama-3.3-70B (Grattafiori et al., 2024), Mistral-Small-24B (Mistral AI, 2025), gpt-oss-120B with 117B total and 5.1B active parameters (OpenAI, 2025), and Qwen3-235B-A22B (235B total, 22B active). Qwen3-8B reads the pages in the one-page-per-origin control of Fig. 2c; under standard retrieval, OriginCount removes 3.7 to 6.1 points of CE from it, 4.90 on average. The default retriever is dense (Karpukhin et al., 2020), and the other two are BM25 (Robertson & Zaragoza, 2009) and an API for web search.

C.3 BASELINES

Each baseline replaces the counting step of the default page-count agent, except for the agents with their own stopping rule, which instead run their own published protocol on the same retrieved pages.

- **Page count.** The default agent of Section 3.1 with the page count (1) and the threshold $\tau = 0.9$.
- **Oracle grouping.** The count (3) on the true groups given by the provenance labels of the snapshots. It uses information that a deployed agent does not have, so it is shown but not ranked.
- **Random grouping.** The read pages are put into random groups at the same retrieval budget and scored with (3); this tests whether any kind of grouping helps or only a grouping by provenance.
- **Verbatim de-duplication.** MinHash over word shingles (Broder, 1997); the survivors are counted.
- **Semantic de-duplication.** Each page whose embedding similarity to a page already kept exceeds a threshold θ is removed (Abbas et al., 2023), and the remaining pages are then counted as in (1).
- **Domain de-duplication.** Each web domain gets a single vote, whatever the number of its pages.
- **Source reliability.** Each page is scaled by the estimated reliability of its site (Hwang et al., 2025).
- **Flat discount.** The confidence of the page-count agent is multiplied by 0.9, so it never reaches τ.
- **Self-reported count.** The backbone reader states how many independent sources support the candidate answer, and this number replaces the page count in the agent (Kadavath et al., 2022).
- **Self-RAG, CRAG, FLARE, Sufficient Context, and Search-o1.** Published retrieval systems that decide for themselves when to retrieve, when to stop, and what to answer (Asai et al., 2024; Yan et al., 2024; Jiang et al., 2023; Joren et al., 2025; Li et al., 2025); they run on the same pages.
- **NLI consistency.** A read page is counted only when a trained entailment model finds it consistent with the candidate answer (Honovich et al., 2022), and the other pages are left out of the count.
- **Citations (ALCE).** Support is counted by means of citation-based attribution (Gao et al., 2023).
- **Conflict weighting.** Support is weighted by conflicts or refutations among pages (Xu et al., 2024).
- **Reliability + de-dup.** Source-reliability weighting after a verbatim de-duplication of the pages.

C.4 IMPLEMENTATION OF ORIGINCOUNT

Algorithm 1 gives the procedure for one question. The pages are taken in reading order, and each page joins the group of an earlier page that it is linked to, trying level 1 first, then level 2, then level 3. Every page thus keeps a link to at most one page read before it, and the groups are the connected components of the kept links, as in Theorem 2. The same groups then serve every candidate answer.

Algorithm 1 OriginCount for one question

Require: question x; read pages $d_1, \dots, d_n$ in reading order; supports $s_a(d_i)$ for every candidate a; an extractor

all links below are read for the claim that x asks about
$\mathcal{P} \leftarrow \emptyset$ ▷ estimated groups
for $i = 1, \dots, n$ **do**
  if d_i and a page of a group $g \in \mathcal{P}$ are linked by a citation, a link, or a quotation **then** add d_i to g ▷ level 1
  else if d_i shares an identifier with a page of a group $g \in \mathcal{P}$ **then** add d_i to g ▷ level 2
  else if the extractor finds that d_i and a page of a group $g \in \mathcal{P}$ restate one another **then** add d_i to g ▷ level 3
  else add the new group $\{d_i\}$ to $\mathcal{P}$ ▷ new origin or untraced page
  end if
end for
for each candidate a **do**
  $\hat{N}(a) \leftarrow \sum_{g \in \mathcal{P}} \frac{1}{|g|} \sum_{d \in g} s_a(d)$
end for
return the counts $\hat{N}(a)$ in place of $N_{\text{page}}(a)$, and keep all pages in the context

Levels 1 and 2 read references and metadata, and level 3 needs an extractor. The default is an LLM extractor, the rule-based extractor uses matching rules only, and the mixed extractor applies the rules first and calls the LLM on part of the pages. Their extra cost is 16.09, 2.02, and 6.94 ms per page. We estimate the wall clock time of a question as 2.80 s of retrieval and reading plus twenty times the extra cost per page, so the LLM extractor adds 0.32 s and the rule-based one 0.04 s (see Table 25).

A worked example. Table 7 follows one question through Algorithm 1. Nine pages are read and seven of them state the candidate answer, so $N_{\text{page}} = 7$, and the page-count agent reports $\sigma(0.35 \times 7) = 0.921$, which is above τ, so it stops. The provenance links give three groups. Group A holds the release d_1 of a statistics office and five pages that copy it, group B holds an independent survey d_5 and one page that quotes it, and the untraced note d_7 forms group C. The count is $\hat{N} = \frac{5}{6} + 1 + 0 = 1.833$, the confidence is $\sigma(0.35 \times 1.833) = 0.655$, and the agent keeps reading instead of stopping. The forum post d_8 restates the release but was read as not supporting. It moves the vote of group A by $-1/6$, the bound $1/(m_g + 1)$ of Theorem 1 at $m_g = 5$, and adds no new vote. The note d_7 keeps its own group and would add a full vote if it stated the candidate answer.

Table 7: Worked example. For each page: its content, its support s, the level of the link that joins it to its group (–: none), the group, its weight w_i, and its contribution $w_i s$.

Page	Content	s	Level	Group	w_i	$w_i s$
d_1	release of the figure by a statistics office	1	–	A	$1/6$	0.167
d_2	news article that cites d_1	1	1	A	$1/6$	0.167
d_3	data portal page with the release number of d_1	1	2	A	$1/6$	0.167
d_4	blog post that paraphrases d_1	1	3	A	$1/6$	0.167
d_5	independent survey that reports the figure	1	–	B	$1/2$	0.500
d_6	page that quotes d_5	1	1	B	$1/2$	0.500
d_7	analyst note, not traced	0	–	C	1	0.000
d_8	forum post that restates d_1	0	3	A	$1/6$	0.000
d_9	translation of d_1	1	3	A	$1/6$	0.167
$N_{\text{page}} = \sum_i s(d_i)$						7
$\hat{N} = \sum_i w_i s(d_i)$						1.833

C.5 Metrics and Intervals

The confident-error rate (CE) is the share of all questions whose answer is wrong and whose confidence is at least 0.8. The expected calibration error (ECE) uses two bins, confident (at least 0.8) and unsure (below 0.8). For each bin it takes the gap between the mean confidence and the accuracy, weights the gap by the share of all questions that fall in the bin, and adds the two terms (Naeini et al., 2015; Guo et al., 2017). Table 8 shows the two bins on NEWS-A for four of the methods. The CE is the mass of the confident bin times its error rate, and the ECE is the sum of the two weighted gaps.

Table 8: Confidence bins on NEWS-A: share of questions (mass), mean confidence, accuracy, and the contribution of each bin to CE and ECE, all as fractions. The remaining mass of 0.025 is the abstentions.

Method	Bin	Mass	Confidence	Accuracy	CE	ECE
Page count	confident	0.414	0.927	0.596	0.167	0.137
	unsure	0.561	0.721	0.733	0	0.007
Oracle grouping	confident	0.330	0.910	0.650	0.116	0.086
	unsure	0.645	0.750	0.712	0	0.025
Self-reported count	confident	0.366	0.803	0.668	0.121	0.049
	unsure	0.609	0.776	0.652	0	0.076
OriginCount	confident	0.320	0.880	0.617	0.123	0.084
	unsure	0.655	0.654	0.704	0	0.033

All comparisons are paired by question. For the six-snapshot means of Table 1, the intervals are 95% confidence intervals of the mean paired difference, with equal weight per snapshot. On a single snapshot the intervals are computed over all its 10,000 questions, for the estimates in Table 11 and for the paired differences in Table 15. All numbers are rounded once, after all computation, so a sum of rounded parts can then differ from the rounded value of the total by a unit or two in the last digit.

C.6 Controlled Experiments

Several analyses use controlled pools, in which the provenance of every page is known by construction. *Added copies* (Fig. 1a and Appendix E.2): starting from a set of pages that holds one supporting origin, we add k pages that copy this origin, or k pages from new origins that support the same answer,

from a site of reliability weight 0.88 or 0.56. *Rewrite strength* (Figs. 1b and 6b): copies rewrite 5%, 15%, 30%, 55%, or 80% of the tokens of their origin. *Verbatim copies* (Fig. 6a): every page of a snapshot is copied verbatim from one to sixteen times. *Noise and missing provenance* (Figs. 8 and 9): each snapshot gives a pool of pages from seven origins, with 24 pages and 276 page pairs in Fig. 9, and the signals that the grouping rules read are perturbed with noise or removed. *Correlated origins* (Fig. 10): five distinct origins are each correct with the accuracy that a single origin has on the snapshot, their correctness follows an exchangeable beta-binomial model with correlation ρ, and the answer is the majority vote, with a confidence of 0.9 on 85% of the questions and of 0.5 on the rest.

D Additional Results

D.1 Accuracy and Stopping Round per Snapshot

Tables 9 and 10 complete Table 1 with the accuracy and stopping round of each method per snapshot. Of the methods that stop, OriginCount stops latest on average, 0.29 rounds later than the page count, since copies no longer raise its confidence early. The oracle stops 0.14 rounds after the page count.

Table 9: Accuracy (%) per snapshot for the methods of Table 1 (N: NEWS, S: SCIENCE, P: PRODUCT).

Method	N-A	N-B	S-A	S-B	P-A	P-B	Mean
Page count (default)	65.81	70.87	76.38	69.92	78.47	76.41	72.98
Oracle grouping	67.39	72.67	74.21	67.50	79.29	76.75	72.97
Random grouping	64.24	65.99	74.68	65.75	74.03	70.17	69.14
Verbatim de-dup	64.12	70.17	73.14	69.22	75.48	80.94	72.18
Semantic de-dup	66.82	68.65	75.86	66.61	76.50	73.62	71.34
Domain de-dup	66.39	65.45	74.70	70.59	78.78	77.20	72.19
Source reliability	64.61	68.50	74.46	68.04	80.03	76.01	71.94
Flat discount	65.50	70.55	76.10	69.70	78.20	75.61	72.61
Self-reported count	64.15	70.64	78.56	69.55	78.58	77.74	73.20
Self-RAG	65.93	71.62	78.21	69.17	76.79	72.98	72.45
CRAG	68.80	66.09	75.32	70.62	76.56	72.18	71.59
FLARE	67.81	65.60	73.40	72.24	76.18	72.48	71.28
Sufficient Context	67.91	70.88	75.59	72.43	74.67	77.42	73.15
Search-o1	69.29	67.21	77.32	70.45	75.45	77.04	72.79
NLI consistency	65.67	71.60	75.84	70.25	76.50	79.72	73.26
Citations (ALCE)	65.44	67.99	72.96	72.03	76.34	73.88	71.44
Conflict weighting	64.71	73.18	73.09	69.51	74.51	73.31	71.38
Reliability + de-dup	66.68	71.65	79.39	71.24	78.72	71.81	73.25
OriginCount (L1–2)	66.14	70.75	75.71	68.60	79.74	79.84	73.46
OriginCount (L1–3)	65.88	69.08	74.65	70.11	77.11	75.10	71.99

D.2 Intervals on NEWS-A

Table 11 gives 95% intervals on NEWS-A, computed over its 10,000 questions, for the page count, the oracle, the self-reported count, reliability with de-duplication, and OriginCount. The CE interval of OriginCount lies below that of the page count and overlaps with those of the other three methods.

D.3 All Methods under Three Retrievers

Table 12 runs every method of Table 1 with BM25, dense, and web-API retrieval. OriginCount has the lowest CE of all methods except the oracle under each retriever. No method, the oracle included, changes by more than 0.75 points of CE between retrievers, and OriginCount changes by 0.27 points.

D.4 All Methods under Five Backbones

Tables 13 and 14 run every method with the five readers. OriginCount has the lowest CE of the methods that can be deployed under every backbone, and its accuracy is between 71.89% and 72.14%.

Table 10: Mean stopping round per snapshot, as defined in Appendix C.1, for the methods of Table 1 (N: NEWS, S: SCIENCE, P: PRODUCT). The flat discount never reaches τ and has no stopping round.

Method	N-A	N-B	S-A	S-B	P-A	P-B	Mean
Page count (default)	3.30	3.00	2.74	3.25	3.83	2.79	3.15
Oracle grouping	3.43	3.10	2.78	3.35	3.98	3.08	3.29
Random grouping	3.59	3.40	2.79	3.33	3.94	2.99	3.34
Verbatim de-dup	3.33	3.33	3.02	3.44	4.18	3.07	3.40
Semantic de-dup	3.53	3.29	2.93	3.48	3.93	3.08	3.37
Domain de-dup	3.33	3.31	2.85	3.53	4.18	2.85	3.34
Source reliability	3.39	3.24	2.87	3.30	3.90	3.03	3.29
Flat discount	–	–	–	–	–	–	–
Self-reported count	3.54	3.10	2.97	3.39	4.14	2.92	3.34
Self-RAG	3.56	3.43	2.96	3.44	4.12	3.00	3.42
CRAG	3.55	3.43	2.83	3.51	4.19	3.06	3.43
FLARE	3.45	3.40	2.97	3.44	4.17	3.02	3.41
Sufficient Context	3.62	3.09	3.03	3.29	4.19	2.88	3.35
Search-o1	3.53	3.14	2.80	3.35	4.13	2.93	3.31
NLI consistency	3.56	3.39	2.80	3.44	4.10	2.98	3.38
Citations (ALCE)	3.50	3.09	3.12	3.53	3.90	3.10	3.37
Conflict weighting	3.60	3.32	2.88	3.57	4.06	2.95	3.40
Reliability + de-dup	3.38	3.21	2.98	3.38	3.89	2.98	3.30
OriginCount (L1–2)	3.39	3.06	3.08	3.33	4.09	3.03	3.33
OriginCount (L1–3)	3.56	3.34	3.06	3.51	4.13	3.05	3.44

Table 11: Point estimates and 95% intervals on NEWS-A (10,000 questions); ECE as a point estimate.

Method	CE (%)	Accuracy (%)	ECE (%)
Page count (default)	16.72 [15.99, 17.45]	65.81 [64.88, 66.74]	14.37
Oracle grouping	11.55 [10.92, 12.18]	67.39 [66.47, 68.31]	11.03
Self-reported count	12.14 [11.50, 12.78]	64.15 [63.21, 65.09]	12.51
Reliability + de-dup	13.48 [12.81, 14.15]	66.68 [65.75, 67.60]	11.08
OriginCount (L1–3)	12.26 [11.62, 12.91]	65.88 [64.95, 66.81]	11.70

D.5 Paired Differences per Snapshot and Backbone

Table 15 gives the paired difference in CE between OriginCount and the page count for all 30 pairs of snapshot and backbone. All 30 intervals lie below zero. The smallest reduction is 3.19 points of CE, for gpt-oss-120B on NEWS-B, and the largest is 6.54 points, for Mistral-Small-24B on SCIENCE-B.

D.6 Retrieval Budget

Fig. 3 varies the number of prefetched calls from 2 to 16 for eight methods. More calls lower the CE of every method on NEWS-A. At 16 calls, OriginCount averages 10.79% over the six snapshots, 0.62 points above the oracle and 0.66 points below the self-reported count, the best of the other methods run at this budget. Its extra cost is the highest of the eight, and Appendix E.6 gives cheaper extractors.

D.7 Calibration

Fig. 4 splits the ECE of the twenty methods. For 15 of them, more than half of the calibration error is made in the confident bin, where confident errors occur. The share is 77% for the page count and 49% for OriginCount. It ranks second of twenty in ECE, behind the flat discount and ahead of the oracle.

D.8 Extra Cost and Retrievers

Fig. 5a shows that the extra work of OriginCount with the LLM extractor takes about a tenth of the estimated time per question, and that of every other method at most 5%. Fig. 5b shows all methods under three retrievers. Fig. 5c places CE against extra cost. Among the methods that can be deployed, the page count, verbatim de-duplication, and the three extractors of OriginCount form the Pareto front, so at any budget of 2.02 ms per page or more, one version of OriginCount has the lowest CE.

Table 12: Six-snapshot CE and accuracy (%) under three retrievers; lowest non-oracle CE per retriever in bold. Range: the largest minus the smallest CE of the method.

	BM25		Dense		Web API		
Method	CE	Acc	CE	Acc	CE	Acc	Range
Page count (default)	16.25	72.91	16.43	72.98	16.08	72.83	0.35
Oracle grouping	11.27	72.97	10.52	72.97	10.85	73.00	0.75
Random grouping	13.89	70.11	14.30	69.14	14.47	69.77	0.58
Verbatim de-dup	13.49	72.53	13.47	72.18	13.54	72.26	0.07
Semantic de-dup	13.19	71.66	12.71	71.34	12.86	71.65	0.48
Domain de-dup	13.78	72.51	13.62	72.19	13.75	72.23	0.16
Source reliability	13.90	72.02	13.96	71.94	14.17	72.13	0.27
Flat discount	15.77	72.66	15.88	72.61	16.04	72.55	0.27
Self-reported count	12.00	73.12	12.31	73.20	12.73	73.19	0.73
Self-RAG	13.46	72.50	13.60	72.45	13.65	72.45	0.19
CRAG	14.20	71.85	14.31	71.59	14.13	71.67	0.18
FLARE	13.76	71.64	13.86	71.28	13.44	71.53	0.42
Sufficient Context	13.06	73.13	12.96	73.15	13.05	73.02	0.10
Search-o1	13.57	72.81	13.05	72.79	13.13	72.74	0.52
NLI consistency	12.61	73.19	12.60	73.26	12.64	73.30	0.04
Citations (ALCE)	12.88	71.68	13.23	71.44	13.13	71.85	0.35
Conflict weighting	13.95	71.54	13.68	71.38	13.22	71.82	0.73
Reliability + de-dup	13.91	73.16	13.75	73.25	13.80	73.10	0.16
OriginCount (L1–2)	13.43	73.31	13.33	73.46	13.46	73.32	0.13
OriginCount (L1–3)	**11.61**	72.03	**11.34**	71.99	**11.39**	71.99	0.27

Table 13: Six-snapshot CE (%) under five backbones (dense retriever). Lowest non-oracle value per column in bold.

Method	Qwen3-32B	Llama-3.3-70B	Mistral-24B	gpt-oss-120B	Qwen3-235B
Page count (default)	16.43	16.08	16.72	15.94	16.61
Oracle grouping	10.52	11.18	11.19	10.76	11.17
Random grouping	14.30	14.00	14.08	14.14	14.14
Verbatim de-dup	13.47	13.65	13.39	13.55	13.21
Semantic de-dup	12.71	12.79	12.90	12.96	13.07
Domain de-dup	13.62	13.51	13.70	13.86	13.59
Source reliability	13.96	13.80	13.94	13.98	13.97
Flat discount	15.88	15.46	16.02	15.31	15.97
Self-reported count	12.31	12.09	12.27	12.53	12.33
Self-RAG	13.60	13.59	13.64	13.59	13.69
CRAG	14.31	13.73	14.06	14.22	14.39
FLARE	13.86	13.46	14.16	13.85	13.75
Sufficient Context	12.96	13.35	12.97	13.02	12.97
Search-o1	13.05	13.66	13.45	12.95	13.18
NLI consistency	12.60	12.75	12.56	12.56	12.85
Citations (ALCE)	13.23	13.42	13.50	13.51	13.51
Conflict weighting	13.68	14.01	13.60	13.99	13.98
Reliability + de-dup	13.75	13.67	13.46	13.75	13.63
OriginCount (L1–2)	13.33	13.61	13.36	13.33	13.20
OriginCount (L1–3)	**11.34**	**11.41**	**10.87**	**12.17**	**10.90**

D.9 Accuracy and Confident Errors

Table 16 splits the accuracy change of OriginCount on each snapshot into the questions moved from wrong to right (R) and from right to wrong (B), next to the change of the oracle. Both groupings raise accuracy on some snapshots and lower it on others. Over the six snapshots the oracle changes mean accuracy by -0.01 points and OriginCount by -0.99 points, which suggests that most of the loss comes from errors of the estimated groups. Table 17 compares OriginCount with the self-reported count per snapshot. OriginCount has the lower CE on four of the six snapshots, with the largest gains on SCIENCE-A and PRODUCT-B, and the self-reported count shows the higher accuracy on four.

Table 14: Six-snapshot accuracy (%) under five backbones (dense retriever).

Method	Qwen3-32B	Llama-3.3-70B	Mistral-24B	gpt-oss-120B	Qwen3-235B
Page count (default)	72.98	73.12	72.75	73.19	72.81
Oracle grouping	72.97	72.82	72.96	72.93	72.93
Random grouping	69.14	69.99	69.68	69.60	69.98
Verbatim de-dup	72.18	72.43	72.31	72.33	72.25
Semantic de-dup	71.34	71.69	71.58	71.68	71.74
Domain de-dup	72.19	72.31	72.28	72.34	72.30
Source reliability	71.94	72.06	72.12	71.98	72.14
Flat discount	72.61	72.78	72.47	72.91	72.52
Self-reported count	73.20	73.02	73.24	73.21	73.11
Self-RAG	72.45	72.55	72.48	72.56	72.50
CRAG	71.59	71.75	71.74	71.88	71.83
FLARE	71.28	71.49	71.56	71.43	71.43
Sufficient Context	73.15	73.15	72.97	72.95	73.02
Search-o1	72.79	72.72	72.79	72.89	72.88
NLI consistency	73.26	73.23	73.24	73.13	73.18
Citations (ALCE)	71.44	71.76	71.67	71.58	71.72
Conflict weighting	71.38	71.73	71.59	71.73	71.68
Reliability + de-dup	73.25	73.20	73.20	73.13	73.29
OriginCount (L1–2)	73.46	73.42	73.43	73.24	73.31
OriginCount (L1–3)	71.99	71.99	71.89	72.14	71.89

Table 15: Paired difference in CE (points), OriginCount minus the page count, with the lower (Low) and upper (High) ends of the 95% interval over 10,000 paired questions.

	NEWS-A			NEWS-B			SCIENCE-A		
Backbone	Diff.	Low	High	Diff.	Low	High	Diff.	Low	High
Qwen3-32B	−4.46	−5.22	−3.79	−4.26	−5.11	−3.53	−5.56	−6.11	−4.93
Llama-3.3-70B	−3.93	−4.70	−3.30	−4.04	−4.92	−3.23	−5.27	−6.04	−4.45
Mistral-24B	−5.60	−6.48	−5.02	−5.31	−6.15	−4.62	−6.43	−7.09	−5.76
gpt-oss-120B	−3.42	−4.12	−2.58	−3.19	−3.80	−2.35	−4.33	−5.12	−3.76
Qwen3-235B	−5.20	−5.86	−4.59	−4.97	−5.61	−4.40	−6.36	−7.22	−5.55
	SCIENCE-B			PRODUCT-A			PRODUCT-B		
Backbone	Diff.	Low	High	Diff.	Low	High	Diff.	Low	High
Qwen3-32B	−5.47	−6.03	−4.67	−5.77	−6.35	−5.07	−5.04	−5.68	−4.41
Llama-3.3-70B	−4.82	−5.39	−4.15	−5.46	−6.07	−4.77	−4.50	−5.30	−3.69
Mistral-24B	−6.54	−7.08	−5.83	−6.25	−7.11	−5.51	−4.97	−5.78	−4.20
gpt-oss-120B	−3.92	−4.68	−3.28	−4.46	−5.03	−3.71	−3.30	−4.12	−2.71
Qwen3-235B	−6.04	−6.84	−5.29	−6.50	−7.09	−5.68	−5.19	−5.90	−4.59

E Additional Ablation Studies

This appendix gives the analyses that support Sections 5.4 and 5.5. Unless stated otherwise, the setting is the default one of Table 1: Qwen3-32B, dense retrieval, and eight prefetched retrieval calls.

E.1 Grouping Quality against the Provenance Labels

Before we rely on the estimated groups, we check them against the labels. Table 18 scores the pairwise decision "same origin" on 2,000 same-origin and 8,000 different-origin page pairs, with labels that trace provenance through levels 1–2 only (top) or through all three levels (bottom). With all three levels, the precision is 0.945 and the false-merge rate is 1.24%, so the grouping seldom joins pages of different origins. Most of its errors are false splits, which move the count towards the page count (Theorem 2(i)), while the rarer false merges can lower it by at most one vote each (Theorem 2(iii)). Table 19 splits the same counts by the type of evidence (news, science, or product).

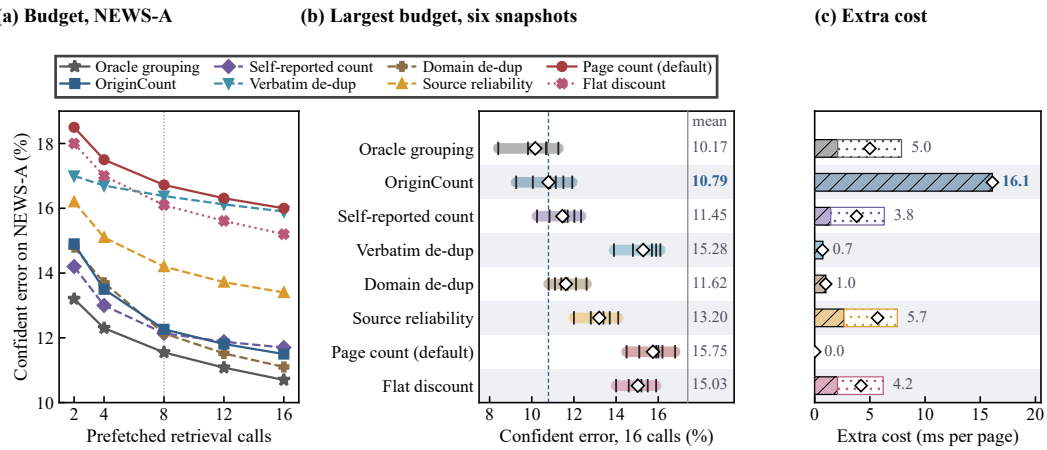


Figure 3: Retrieval budget and extra cost. (a) CE on NEWS-A for eight methods at 2 to 16 prefetched calls; the dotted line marks the default budget of eight calls. (b) CE at 16 calls on the six snapshots: bands span the snapshots, ticks mark single snapshots, diamonds mark the means listed on the right, and the dashed line marks the mean of OriginCount. (c) Extra cost per page: hatched bars reach the lowest value over the snapshots, dotted bars the highest, and diamonds with labels mark the six-snapshot mean of Table 1.

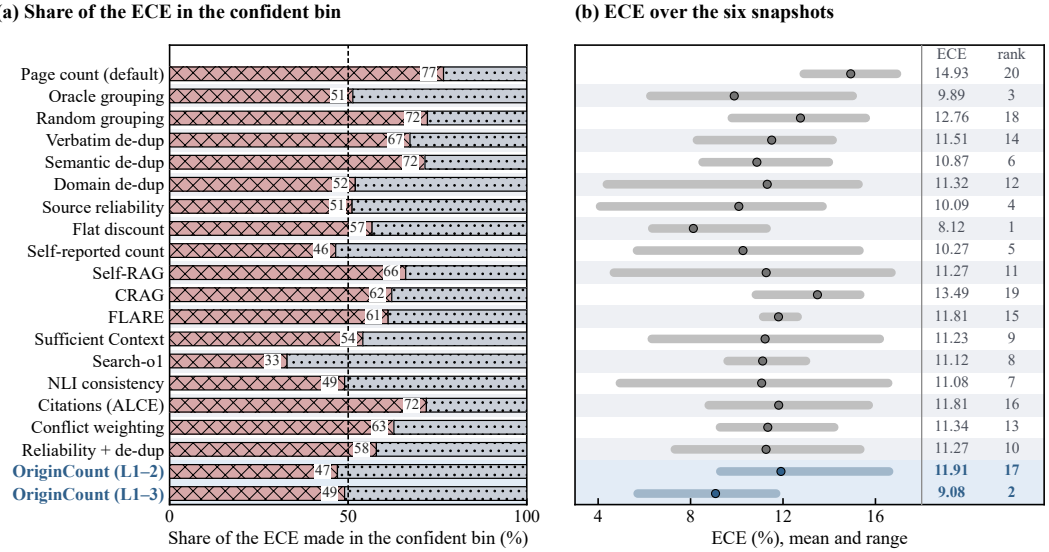


Figure 4: Calibration of the twenty methods. (a) Share of the ECE made in the confident bin (confidence at least 0.8), six snapshots pooled; the dashed line marks one half. (b) ECE over the six snapshots: mean (dot) and range (band), with the six-snapshot mean of Table 1 and the rank among the twenty methods on the right. L1–2 and L1–3 as in Table 1.

E.2 Controlled Copies

Table 20 crosses the reliability weight of the site of four added pages, 0.88 or 0.56, with whether these pages copy the origin already in the set or come from new origins. The page count moves by the same amount in all four cells, and reliability weighting moves by an amount set by the weight in both arms. Only the origin count separates the two kinds of page: it does not move for copies and moves as much as the page count for new origins. Table 21 extends the copy arm from zero to eight pages.

E.3 Decomposition of the CE Reduction

Table 22 splits the CE reduction of OriginCount on each snapshot into three cumulative parts, one per design choice of Section 4.1: grouping the pages by provenance, giving each group one vote, and keeping untraced pages as their own groups. Each part lowers CE on every snapshot, and one vote per group gives the largest part on five of the six. Table 23 lists the confident errors of Fig. 2a by cause.

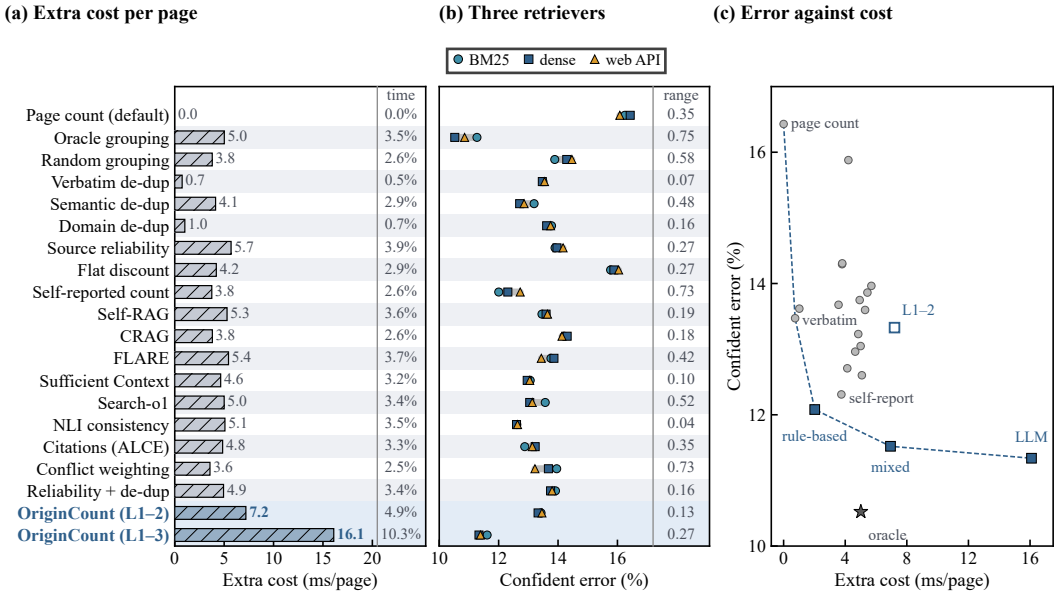


Figure 5: Extra cost and retrievers. (a) Extra cost per page, with its share of the estimated time per question on the right. (b) CE under BM25, dense, and web-API retrieval; the band spans the three retrievers, and the right column gives the range. (c) CE against extra cost. Grey dots: baselines; star: oracle; filled squares: OriginCount with levels 1–3 and the LLM, mixed, and rule-based extractors; hollow square: levels 1–2; dashed line: the Pareto front of the methods that can be deployed.

Table 16: Accuracy (%) of the page count, the oracle, and OriginCount, the change against the page count, and the shares R and B of questions moved by OriginCount (%).

		Oracle grouping		OriginCount				
Snapshot	Page count	Acc	Change	Acc	Change	R	B	$R-B$
NEWS-A	65.81	67.39	+1.58	65.88	+0.07	4.76	4.69	+0.07
NEWS-B	70.87	72.67	+1.80	69.08	−1.79	3.14	4.93	−1.79
SCIENCE-A	76.38	74.21	−2.17	74.65	−1.73	4.41	6.13	−1.72
SCIENCE-B	69.92	67.50	−2.42	70.11	+0.19	4.49	4.30	+0.19
PRODUCT-A	78.47	79.29	+0.82	77.11	−1.36	2.94	4.30	−1.36
PRODUCT-B	76.41	76.75	+0.34	75.10	−1.31	4.22	5.53	−1.31
Mean	72.98	72.97	−0.01	71.99	−0.99	3.99	4.98	−0.99

E.4 De-duplication

Fig. 6 shows the two kinds of de-duplication and the pages per origin that make copies matter. After verbatim de-duplication, CE moves by at most 0.16 points as every page is copied up to sixteen times. Semantic de-duplication lets rewritten copies through, and the CE after it rises as copies are rewritten and levels off after a 30% rewrite, which reaches 96% to 99% of the full rise. Over 50 sampled questions per snapshot, most questions see four or more pages per origin, with a mean of 4.36 to 4.91.

E.5 Untraced Pages

Fig. 7 compares three policies for the 6.5% to 7.7% of pages that no level can link. Keeping each of them as its own group gives the highest accuracy and the lowest CE on every snapshot. Dropping them costs 0.43 to 0.62 points of accuracy and 0.41 to 0.55 points of CE. Attaching them to the nearest group costs 0.69 to 0.91 points of accuracy and 1.16 to 1.69 points of CE, about 0.2 points of CE per point of untraced share, and between 2.2 and 3.1 times the CE cost of dropping them. Table 24 hides provenance from a growing share of the pages. The reduction of CE grows with the traced share and is zero when nothing can be traced, where OriginCount gives the output of the page-count agent.

Table 17: OriginCount and the self-reported count per snapshot: CE and accuracy (%), and the difference, OriginCount minus the self-reported count.

	Self-reported count		OriginCount		Difference	
Snapshot	CE	Acc	CE	Acc	CE	Acc
NEWS-A	12.14	64.15	12.26	65.88	+0.12	+1.73
NEWS-B	12.58	70.64	12.18	69.08	−0.40	−1.56
SCIENCE-A	16.32	78.56	13.23	74.65	−3.09	−3.91
SCIENCE-B	13.38	69.55	13.24	70.11	−0.14	+0.56
PRODUCT-A	7.69	78.58	7.94	77.11	+0.25	−1.47
PRODUCT-B	11.77	77.74	9.18	75.10	−2.59	−2.64
Mean	12.31	73.20	11.34	71.99	−0.97	−1.21

Table 18: Pairwise grouping quality: precision, recall, mean absolute error of the number of estimated groups $\hat{G}$, false-merge rate among different-origin pairs, and false-split rate among same-origin pairs, with labels that trace provenance through levels 1–2 (top) or all three levels (bottom).

Labels	Levels used	Precision	Recall	MAE of $\hat{G}$	False merge	False split
Levels 1–2	1	0.983	0.795	0.284	0.35%	20.6%
	1–2	0.961	0.886	0.294	0.90%	11.5%
	1–3	0.955	0.912	0.238	1.08%	8.9%
Levels 1–3	1	0.951	0.830	0.389	1.08%	17.0%
	1–2	0.945	0.847	0.157	1.23%	15.3%
	1–3	0.945	0.854	0.197	1.24%	14.6%

E.6 Provenance Levels, Extractors, and Cost

Table 25 asks whether a cheaper configuration can replace the full grouping. The rule-based extractor keeps CE at 12.08%, below every baseline in Table 1, at one eighth of the extra cost and 0.04 s of extra wall clock per question. The mixed extractor comes within 0.18 points of the LLM extractor at under half its cost. Levels 1–2 give the highest accuracy, and level 1 alone keeps less of the reduction.

E.7 Noise in the Grouping Signals

Fig. 8 adds noise to the signals of three grouping rules on controlled pools with seven origins per snapshot. The provenance rule of OriginCount stays within one origin of the truth on every snapshot and at every noise level. Embedding clustering overcounts more as the noise grows, by 6.3 origins on average at the highest level, and a hybrid of the two lies in between once noise is added. For both similarity-based rules, wrong splits and merges grow with the noise, and splits outnumber merges.

E.8 Partially Observed Provenance

Fig. 9 removes the provenance of a growing share of pages in controlled pools. With all links observed, the provenance rule groups every page pair correctly, the domain rule makes wrong splits and wrong merges on every snapshot, and a rule that uses only citations makes a few wrong merges and undercounts by one or two origins. As provenance is removed from 38% and 75% of the pages, the provenance rule overcounts but does not undercount, since it makes no wrong merge at any untraced share. This agrees with Theorem 2(i): with true links only, the count stays at or above N_{orig}.

E.9 Copies with New Facts and Discarded Content

Assumption 1 holds for a copy that adds nothing to its origin. Table 26 breaks it: some copies carry a new observation and are merged anyway. Accuracy falls by 0.14 to 0.23 points when a tenth of the copies carry a new fact and by 1.04 to 1.31 points when half of them do, while CE rises by no more than 0.94 points. Table 27 also discards the clarifying content that grouped pages add, which raises CE by 0.48 points on average. OriginCount keeps all pages in the context and thus avoids this rise.

Table 19: Pairwise counts by type of evidence: true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN) of the decision "same origin", with the rates of Table 18 as fractions.

Levels	Type	TP	FN	FP	TN	Precision	Recall	False merge	False split
Labels through levels 1–2									
1	news	526	140	9	2710	0.983	0.790	0.0033	0.210
1	science	562	123	12	2580	0.979	0.820	0.0046	0.180
1	product	501	148	7	2682	0.986	0.772	0.0026	0.228
1–2	news	581	85	21	2698	0.965	0.872	0.0077	0.128
1–2	science	616	69	23	2569	0.964	0.899	0.0089	0.101
1–2	product	574	75	28	2661	0.953	0.884	0.0104	0.116
1–3	news	606	60	29	2690	0.954	0.910	0.0107	0.090
1–3	science	625	60	30	2562	0.954	0.912	0.0116	0.088
1–3	product	592	57	27	2662	0.956	0.912	0.0100	0.088
Labels through all three levels									
1	news	542	118	28	2635	0.951	0.821	0.0105	0.179
1	science	562	116	35	2613	0.941	0.829	0.0132	0.171
1	product	556	106	23	2666	0.960	0.840	0.0086	0.160
1–2	news	559	101	31	2632	0.947	0.847	0.0116	0.153
1–2	science	565	113	33	2615	0.945	0.833	0.0125	0.167
1–2	product	570	92	34	2655	0.944	0.861	0.0126	0.139
1–3	news	556	104	33	2630	0.944	0.842	0.0124	0.158
1–3	science	586	92	26	2622	0.958	0.864	0.0098	0.136
1–3	product	566	96	40	2649	0.934	0.855	0.0149	0.145

Table 20: Four added pages from a site of reliability weight 0.88 or 0.56, added as copies of the origin in the set ($G = 1$) or as new origins ($G = 5$): change of confidence and of the expected stopping round under the page count, reliability weighting, and the origin count.

		Δ confidence			Δ stopping round		
Weight	Added pages	Page	Reliability	Origin	Page	Reliability	Origin
0.88	copies ($G = 1$)	$+0.265$	$+0.247$	0.000	-0.53	-0.49	0.00
0.88	new origins ($G = 5$)	$+0.265$	$+0.247$	$+0.265$	-0.53	-0.49	-0.53
0.56	copies ($G = 1$)	$+0.265$	$+0.178$	0.000	-0.53	-0.36	0.00
0.56	new origins ($G = 5$)	$+0.265$	$+0.178$	$+0.265$	-0.53	-0.36	-0.53

E.10 Combination with Other Agents

Table 28 gives the combination results of Table 2 together with the interaction and a re-run of each agent, and Table 29 gives the CE of OriginCount per agent and snapshot. For every agent, the CE removed by the two changes together is within 0.10 points of the sum of the two separate gains, and the re-runs reproduce the reported value of each agent within 0.11 points. The logistic gate gains least from OriginCount, 2.82 points, and the two counting variants gain the most, 5.09 points each.

E.11 Experiments with Correlated Origins

Fig. 10 measures the dependence studied in Appendix B.5 on controlled pools of five distinct origins. At $\rho = 0$ the majority of five origins is 13 to 15 points more accurate than a single origin. As ρ grows, the effective number of origins falls to 1.19 at $\rho = 0.8$, the gain of the majority vote nearly vanishes, and CE rises by 11.1 to 12.2 points at $\rho = 0.8$. On NEWS-A, SCIENCE-A, and PRODUCT-A, CE rises from 15.01%, 12.76%, and 10.69% at $\rho = 0$ to 24.82%, 23.02%, and 21.24% at $\rho = 0.5$. Grouping by provenance cannot see this kind of correlation, because the correlated origins are distinct.

F Comparison of Counting Units

Table 30 places the main counting units of the paper under the confidence map of our agents. The limit column gives the confidence reached as the number of rewritten copies of one origin, spread over many sites, grows without bound. Only the origin unit, with the copies traced, has a limit set by the evidence instead of by the number of copies, and this limit is the confidence given by one observation. The flat discount caps the confidence at 0.9 but approaches it under these same copies.

Table 21: Confidence after k added pages. In the copy arms the true origin count stays at 1, and in the new-origin arm it is $1 + k$.

	Copies, weight 0.88			Copies, weight 0.56			New origins
k	Page	Reliability	Origin	Page	Reliability	Origin	Origin
0	0.587	0.576	0.587	0.587	0.549	0.587	0.587
2	0.741	0.716	0.587	0.741	0.643	0.587	0.741
4	0.852	0.823	0.587	0.852	0.727	0.587	0.852
6	0.921	0.896	0.587	0.921	0.798	0.587	0.921
8	0.959	0.941	0.587	0.959	0.854	0.587	0.959

Table 22: The CE reduction of OriginCount against the page count (points), split into three cumulative parts.

Snapshot	Grouping	One vote per group	Untraced kept	Total
NEWS-A	1.10	2.95	0.41	4.46
NEWS-B	2.73	1.01	0.52	4.26
SCIENCE-A	1.74	3.30	0.52	5.56
SCIENCE-B	1.85	3.07	0.55	5.47
PRODUCT-A	1.49	3.73	0.55	5.77
PRODUCT-B	1.26	3.32	0.46	5.04
Mean	1.70	2.90	0.50	5.09

Table 23: Confident errors (% of questions) by cause, for the page count (before) and OriginCount (after). Retrieval: no read page states the correct answer; the other causes are errors of extraction, of aggregation, and of other kinds. These four causes are listed once for both counts.

	Stopped early		Wrong consensus		Other causes				Total	
Snapshot	before	after	before	after	Retrieval	Extraction	Aggregation	Other	before	after
NEWS-A	6.34	3.85	4.86	2.90	2.82	1.11	0.93	0.66	16.72	12.26
NEWS-B	6.11	3.51	4.90	3.24	0.68	2.27	1.01	1.47	16.44	12.18
SCIENCE-A	7.17	4.21	5.43	2.81	2.74	0.57	1.69	1.21	18.79	13.23
SCIENCE-B	7.06	3.89	5.48	3.18	0.14	2.54	1.37	2.14	18.71	13.24
PRODUCT-A	5.33	1.64	3.85	1.78	0.22	1.35	0.79	2.17	13.71	7.94
PRODUCT-B	5.56	2.78	3.98	1.70	1.98	1.01	0.97	0.73	14.22	9.18

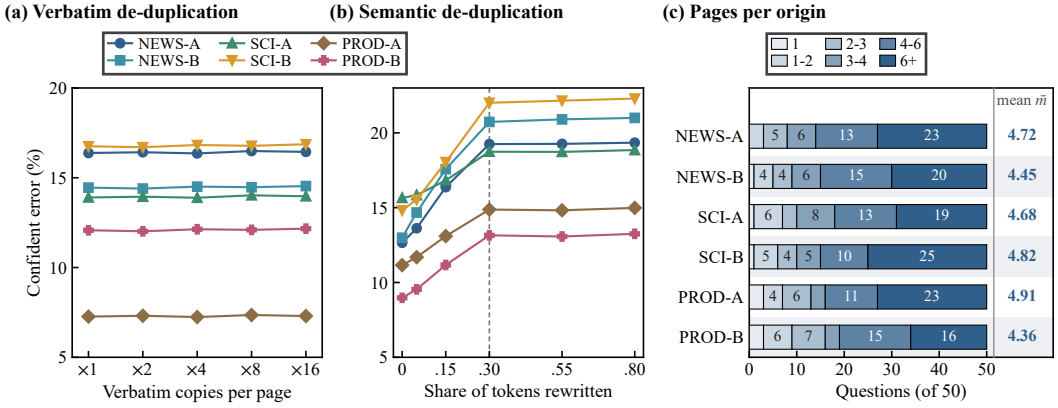


Figure 6: De-duplication and pages per origin. (a) CE after verbatim de-duplication when every page of a snapshot is copied verbatim one to sixteen times. (b) CE after semantic de-duplication when copies rewrite a growing share of their tokens; the dashed line marks a 30% rewrite. (c) Pages per origin $\bar{m}$ of 50 sampled questions per snapshot, binned, with the mean on the right. SCI and PROD abbreviate SCIENCE and PRODUCT.

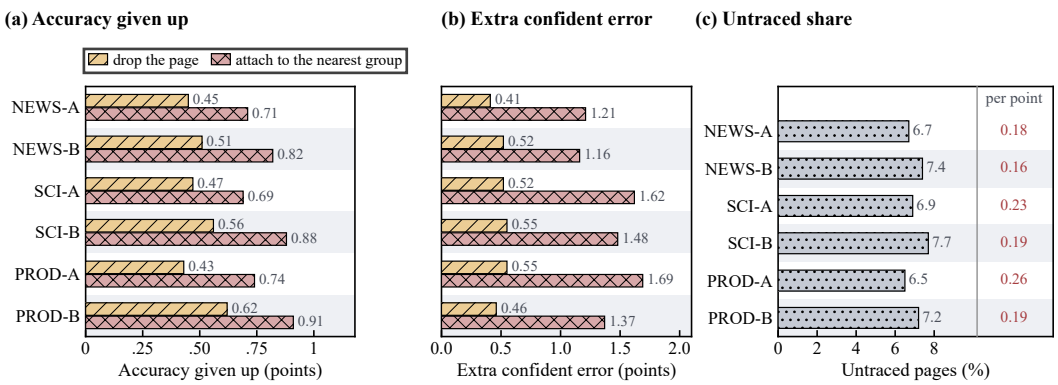


Figure 7: Policies for untraced pages, against keeping each untraced page as its own group. (a) Accuracy given up by dropping untraced pages and by attaching them to the nearest group. (b) Extra CE of the same two policies. (c) Share of untraced pages; right: extra CE of attaching per point of untraced share. SCI and PROD abbreviate SCIENCE and PRODUCT.

Table 24: Provenance hidden from a growing share of pages: six-snapshot CE of OriginCount (%), its change against the page count (points), and the abstention rate (%).

Untraced pages	7% (default)	40%	75%	100%
CE	11.34	13.15	15.06	16.43
ΔCE	−5.09	−3.29	−1.37	0.00
Abstain	2.5	2.7	2.4	2.5

Table 25: Cheaper configurations: extra cost per page, estimated wall clock per question, untraced share, and six-snapshot CE, accuracy, and stopping round.

Configuration	ms/page	Wall clock (s)	Untraced	CE ↓	Acc ↑	Round
Levels 1–3, LLM extractor	16.09	3.12	7.0%	11.34	71.99	3.44
Levels 1–3, mixed extractor	6.94	2.94	7.3%	11.52	71.84	3.46
Levels 1–3, rule-based extractor	2.02	2.84	8.1%	12.08	71.37	3.52
Levels 1–2	7.20	2.94	7.0%	13.33	73.46	3.33
Level 1 only	6.10	2.92	7.8%	13.88	72.06	3.36
Verbatim de-dup only	0.75	2.81	7.0%	13.47	72.18	3.40

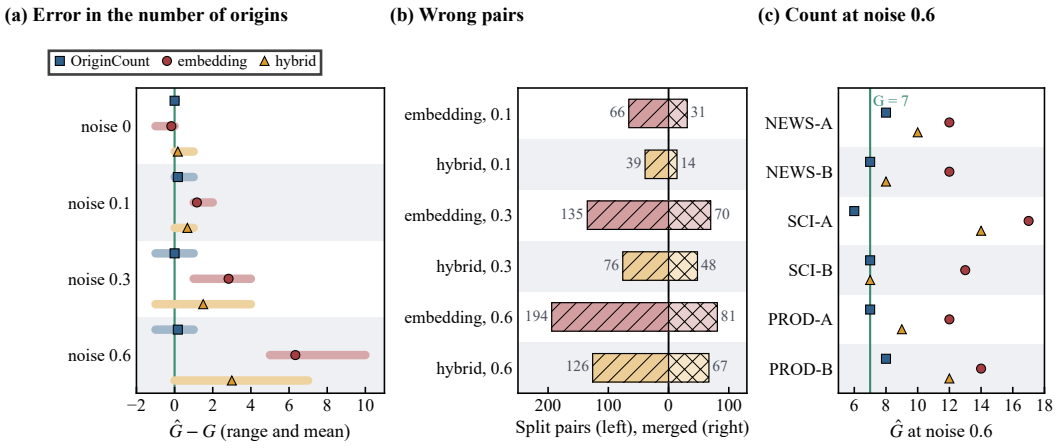


Figure 8: Noise in the grouping signals (controlled pools, $G = 7$ per snapshot). (a) Signed error $\hat{G} - G$, with $\hat{G}$ the number of estimated groups, of three grouping rules at four noise levels: range over the six snapshots (band) and mean (marker). (b) Wrongly split and wrongly merged page pairs of the embedding and hybrid rules, six snapshots pooled. (c) $\hat{G}$ per snapshot at noise 0.6; the green line marks $G = 7$. SCI and PROD abbreviate SCIENCE and PRODUCT.

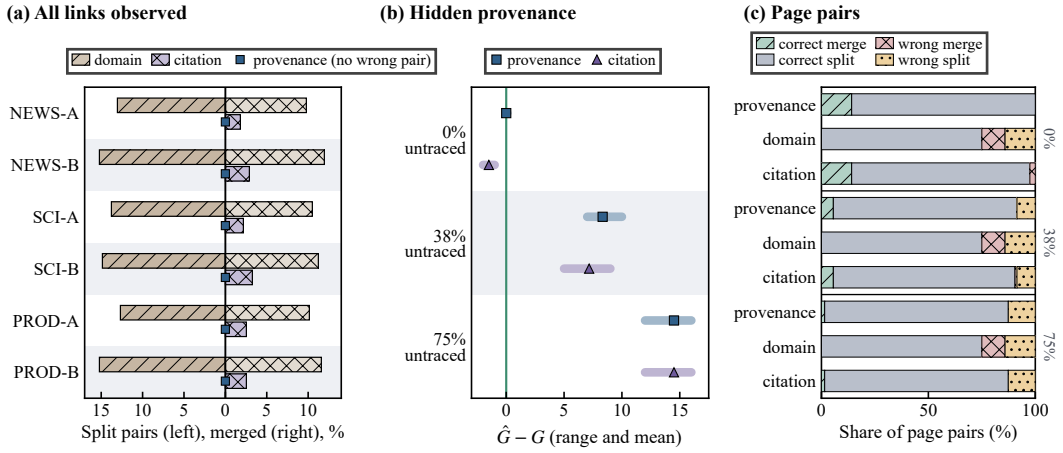


Figure 9: Partially observed provenance (controlled pools with 24 pages and 7 origins per snapshot, 276 page pairs each). (a) Wrongly split (left) and wrongly merged (right) pairs of the domain and citation rules with all links observed, as a share of the pairs; the provenance rule makes no wrong pair. (b) Signed error $\hat{G} - G$, with $\hat{G}$ the number of estimated groups, of the provenance and citation rules with provenance removed from 0%, 38%, and 75% of the pages: range over the snapshots and mean. (c) Composition of the 1,656 page pairs of the six snapshots for the three rules at three untraced shares; the side labels give the share of pages without provenance. SCI and PROD abbreviate SCIENCE and PRODUCT.

Table 26: Copies that carry a new observation and are merged anyway: accuracy, accuracy loss, and CE (%).

	NEWS-A			SCIENCE-A			PRODUCT-A		
Share with new facts	Acc	Loss	CE	Acc	Loss	CE	Acc	Loss	CE
0%	65.88	0.00	12.26	74.65	0.00	13.23	77.11	0.00	7.94
10%	65.71	0.17	12.35	74.42	0.23	13.36	76.97	0.14	8.02
25%	65.36	0.52	12.56	74.04	0.61	13.61	76.64	0.47	8.22
50%	64.72	1.16	13.20	73.34	1.31	14.16	76.07	1.04	8.88

Table 27: Stress settings on NEWS-A, SCIENCE-A, and PRODUCT-A, CE (%). Both settings change the data or the context, not the counting rule.

Setting	NEWS-A	SCIENCE-A	PRODUCT-A	Mean
OriginCount (L1–3)	12.26	13.23	7.94	11.14
Half of the copies carry new facts	13.20	14.16	8.88	12.08
Clarifying content of grouped pages discarded	12.74	13.70	8.42	11.62

Table 28: Change of CE (points) against the page-count agent from each agent's own change (Own), from OriginCount (OC), and from both, with their sum before rounding, the interaction (sum minus both), and the difference between a re-run of the agent and its reported value.

Agent variant	Own	OC	Both	Sum	Interaction	Re-run
Answer voting	−1.82	−5.09	−6.89	−6.91	−0.02	−0.06
Page threshold	−1.73	−5.09	−6.73	−6.83	−0.10	−0.02
FLARE	−1.82	−4.58	−6.43	−6.40	+0.03	−0.04
Sufficient Context	−1.63	−4.58	−6.17	−6.21	−0.04	+0.11
Logistic gate	−1.90	−2.82	−4.62	−4.72	−0.10	+0.10

Table 29: CE (%) of the page-count agent and of each agent variant with OriginCount but without its own change, per snapshot (N: NEWS, S: SCIENCE, P: PRODUCT).

Agent variant	N-A	N-B	S-A	S-B	P-A	P-B	Mean
Page-count agent (all variants)	16.72	16.44	18.79	18.71	13.71	14.22	16.43
Page threshold, answer voting	12.26	12.18	13.23	13.24	7.94	9.18	11.34
Sufficient Context, FLARE	12.55	12.82	13.65	13.91	8.44	9.73	11.85
Logistic gate	14.44	13.96	15.60	15.59	10.60	11.48	13.61

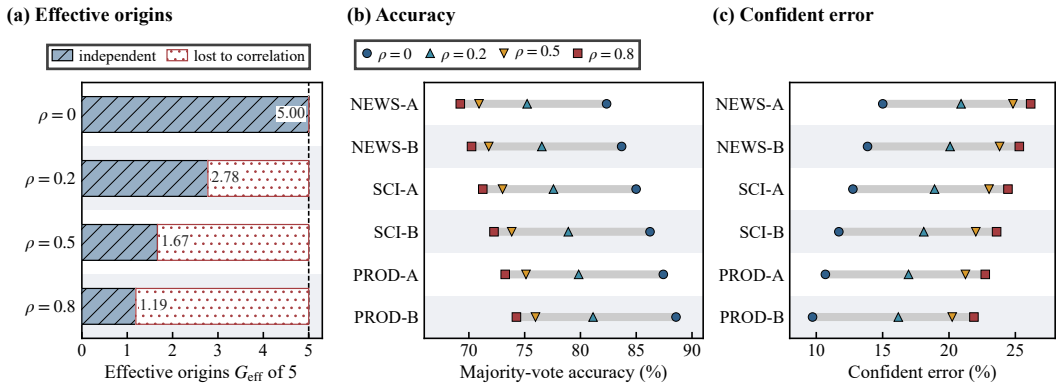


Figure 10: Correlated origins (controlled pools of five distinct origins). (a) Effective number of independent origins $G_{\text{eff}} = 5/(1 + 4\rho)$, with ρ the correlation between origins. (b) Majority-vote accuracy per snapshot at four correlations. (c) CE per snapshot at the same correlations. SCI and PROD abbreviate SCIENCE and PRODUCT.

Table 30: Counting units under the confidence $f(N) = \sigma(0.35N)$: what each unit counts, the limit of the confidence as rewritten copies of one origin accumulate, and the information each unit needs.

Unit	Count N	Limit	Information needed
Page	pages	1	none
Verbatim de-dup	surviving pages	1	hashes
Semantic de-dup	surviving pages	1	embeddings, threshold
Domain	domains	1	domain names
Reliability weighting	$\sum_i q_i$	1	site weights
Square-root weights	pages, weight $\lvert S_g\rvert^{-1/2}$	1	provenance
Flat discount	pages, confidence $\times$ 0.9	0.9	one constant
NLI consistency	consistent pages	1	entailment model
Origin group	origins	$\sigma(0.35) = 0.587$	provenance

G VALUES PLOTTED IN THE FIGURES

Tables 31 to 41 list values plotted in the figures. All values come from the same runs as the tables above and are rounded for display, so a value that is shown in two places can differ in the last digit.

Table 31: Share of rewritten copies of one origin that survive de-duplication (Fig. 1b), by the share of tokens rewritten.

Tokens rewritten	5%	15%	30%	55%	80%
Verbatim (MinHash)	0.33	0.41	0.58	0.69	0.82
Semantic (embedding)	0.14	0.26	0.28	0.32	0.35

Table 32: Copies among the retrieved pages and questions whose stopping round or answer changes when the agent counts origins with OriginCount (Fig. 1c), in %, and the share of the changed questions in which only the stopping round changes.

Snapshot	Copies	Changed	Round only
NEWS-A	81.84	17.50	40.93
NEWS-B	80.35	18.74	50.93
SCIENCE-A	80.49	20.23	43.89
SCIENCE-B	81.92	17.03	41.61
PRODUCT-A	80.56	16.51	50.09
PRODUCT-B	78.35	17.99	39.23

Table 33: CE removed by OriginCount (points) with three backbones, and pages per origin $\bar{m}$ among the pages read (Fig. 2c). In the Qwen3-8B control every origin contributes one page.

	CE removed			Pages per origin $\bar{m}$	
Snapshot	Qwen3-32B	Llama-3.3-70B	Qwen3-8B control	Qwen3-32B	Llama-3.3-70B
NEWS-A	4.46	3.93	0.00	5.5	5.3
NEWS-B	4.26	4.04	0.00	5.1	4.9
SCIENCE-A	5.56	5.27	0.00	5.1	5.3
SCIENCE-B	5.47	4.82	0.00	5.5	5.4
PRODUCT-A	5.77	5.46	0.00	5.1	5.0
PRODUCT-B	5.04	4.50	0.00	4.6	4.8

Table 34: CE (%) after verbatim de-duplication when every page is copied verbatim one to sixteen times (left, Fig. 6a), and after semantic de-duplication when copies rewrite a share of their tokens (right, Fig. 6b).

	Verbatim copies per page					Tokens rewritten				
Snapshot	×1	×2	×4	×8	×16	5%	15%	30%	55%	80%
NEWS-A	16.38	16.42	16.36	16.49	16.44	13.62	16.37	19.25	19.27	19.35
NEWS-B	14.45	14.40	14.51	14.47	14.54	14.68	17.57	20.73	20.90	21.00
SCIENCE-A	13.90	13.94	13.89	14.02	13.97	15.84	16.83	18.74	18.73	18.85
SCIENCE-B	16.75	16.70	16.82	16.78	16.87	15.54	18.05	22.01	22.15	22.29
PRODUCT-A	7.27	7.31	7.24	7.36	7.30	11.69	13.10	14.87	14.82	14.98
PRODUCT-B	12.08	12.02	12.13	12.10	12.17	9.54	11.17	13.15	13.07	13.25

Table 35: Pages per origin $\bar{m}$ of 50 sampled questions per snapshot (Fig. 6c): the number of questions in each range of $\bar{m}$, the smallest, median, and largest numbers of pages and of origins per question, and the mean $\bar{m}$.

	Questions with $\bar{m}$ in						Pages	Origins	
Snapshot	1	1–2	2–3	3–4	4–6	6+	min/med/max	min/med/max	Mean $\bar{m}$
NEWS-A	0	3	5	6	13	23	3/18/20	1/4/10	4.72
NEWS-B	1	4	4	6	15	20	4/17/20	1/4/9	4.45
SCIENCE-A	1	6	3	8	13	19	2/18/20	1/4/10	4.68
SCIENCE-B	1	5	4	5	10	25	4/18/20	1/4/9	4.82
PRODUCT-A	3	4	6	3	11	23	4/19/20	1/4/10	4.91
PRODUCT-B	3	6	7	3	15	16	2/17/20	1/4/9	4.36

Table 36: CE (%) of eight methods on NEWS-A at five retrieval budgets (left, Fig. 3a) and at 16 prefetched calls on the six snapshots (right, Fig. 3b), rounded to one decimal, with the six-snapshot mean.

	NEWS-A, prefetched calls					16 prefetched calls						
Method	2	4	8	12	16	N-A	N-B	S-A	S-B	P-A	P-B	Mean
Oracle grouping	13.2	12.3	11.5	11.1	10.7	10.7	11.3	10.1	10.7	9.8	8.4	10.17
Self-reported count	14.2	13.0	12.1	11.9	11.7	11.7	12.3	11.6	12.0	10.8	10.3	11.45
Verbatim de-dup	17.0	16.7	16.4	16.1	15.9	15.9	16.1	15.3	15.7	14.8	13.9	15.28
Domain de-dup	14.8	13.7	12.1	11.5	11.1	11.1	12.6	11.7	12.1	11.4	10.8	11.62
Source reliability	16.2	15.1	14.2	13.7	13.4	13.4	14.1	13.2	13.7	12.8	12.0	13.20
Page count (default)	18.5	17.5	16.7	16.3	16.0	16.0	16.8	15.9	16.2	15.1	14.5	15.75
Flat discount	18.0	17.0	16.1	15.6	15.2	15.2	15.9	15.0	15.5	14.6	14.0	15.03
OriginCount	14.9	13.5	12.3	11.8	11.5	11.5	11.9	10.9	11.1	10.1	9.3	10.79

Table 37: Calibration of the twenty methods (Fig. 4): share of the ECE made in the confident bin (%), six-snapshot mean ECE (%) with its smallest and largest value over the snapshots, and the rank by mean ECE.

Method	Share in confident bin	ECE over the six snapshots			Rank
		Mean	Min	Max	
Page count (default)	76.7	14.93	12.90	16.93	20
Oracle grouping	51.3	9.89	6.26	15.02	3
Random grouping	72.2	12.76	9.78	15.58	18
Verbatim de-dup	67.3	11.51	8.27	14.15	14
Semantic de-dup	71.5	10.87	8.52	13.98	6
Domain de-dup	52.0	11.32	4.38	15.27	12
Source reliability	51.1	10.09	4.08	13.71	4
Flat discount	56.7	8.12	6.34	11.29	1
Self-reported count	46.5	10.27	5.68	15.31	5
Self-RAG	66.1	11.27	4.68	16.69	11
CRAG	62.2	13.49	10.82	15.34	19
FLARE	61.2	11.81	11.14	12.63	15
Sufficient Context	54.1	11.23	6.32	16.19	9
Search-o1	32.9	11.12	9.60	13.00	8
NLI consistency	49.0	11.08	4.94	16.55	7
Citations (ALCE)	71.9	11.81	8.79	15.71	16
Conflict weighting	62.8	11.34	9.27	14.22	13
Reliability + de-dup	57.9	11.27	7.32	15.34	10
OriginCount (L1–2)	47.0	11.91	9.29	16.59	17
OriginCount (L1–3)	49.0	9.08	5.71	11.70	2

Table 38: Estimated number of origins $\hat{G}$ per snapshot on the controlled pools with $G = 7$ (Fig. 8), for three grouping rules at four noise levels, and the number of wrong page pairs of each rule at noise 0.6.

Rule	Noise	N-A	N-B	S-A	S-B	P-A	P-B
OriginCount	0	7	7	7	7	7	7
	0.1	7	7	7	8	7	7
	0.3	7	8	7	7	6	7
	0.6	8	7	6	7	7	8
Embedding	0	7	7	7	7	7	6
	0.1	8	8	8	9	8	8
	0.3	11	10	8	10	9	11
	0.6	12	12	17	13	12	14
Hybrid	0	7	7	8	7	7	7
	0.1	8	8	7	7	8	8
	0.3	11	9	6	7	8	10
	0.6	10	8	14	7	9	12
Wrong page pairs at noise 0.6							
OriginCount		5	3	4	5	4	6
Embedding		48	60	44	44	37	42
Hybrid		33	36	31	29	32	32

Table 39: Wrongly split and wrongly merged page pairs of the embedding and hybrid rules on the controlled pools, six snapshots pooled (Fig. 8b).

Rule	Noise 0.1		Noise 0.3		Noise 0.6	
	Split	Merged	Split	Merged	Split	Merged
Embedding	66	31	135	70	194	81
Hybrid	39	14	76	48	126	67

Table 40: Partially observed provenance on the controlled pools with $G = 7$ (Fig. 9). Top: signed error $\hat{G} - G$ of the provenance and citation rules per snapshot. Middle: wrong pairs of each rule with all links observed, out of 276 pairs per snapshot. Bottom: the 1,656 page pairs of the six snapshots by outcome.

Untraced pages	Rule	N-A	N-B	S-A	S-B	P-A	P-B
0%	provenance	0	0	0	0	0	0
	citation	−2	−1	−2	−1	−2	−1
38%	provenance	+7	+9	+9	+10	+7	+8
	citation	+6	+9	+8	+9	+5	+6
75%	provenance	+16	+15	+15	+15	+14	+12
	citation	+16	+15	+15	+15	+14	+12
0%	domain, split	36	42	38	41	35	42
	domain, merged	27	33	29	31	28	32
	citation, merged	5	8	6	9	7	7
	provenance, any	0	0	0	0	0	0

Untraced pages	Rule	Right merge	Right split	Wrong merge	Wrong split
0%	provenance	234	1422	0	0
	domain	0	1242	180	234
	citation	234	1380	42	0
38%	provenance	92	1422	0	142
	domain	0	1242	180	234
	citation	92	1407	15	142
75%	provenance	25	1422	0	209
	domain	0	1242	180	234
	citation	25	1422	0	209

Table 41: Untraced pages per snapshot (Fig. 7): share of untraced pages (%), accuracy (%) when each untraced page is its own group, and the accuracy given up and the extra CE (points) from dropping untraced pages or attaching them to the nearest group.

			Accuracy given up		Extra CE	
Snapshot	Untraced	Accuracy	Drop	Nearest	Drop	Nearest
NEWS-A	6.7	65.88	0.45	0.71	0.41	1.21
NEWS-B	7.4	69.08	0.51	0.82	0.52	1.16
SCIENCE-A	6.9	74.65	0.47	0.69	0.52	1.62
SCIENCE-B	7.7	70.11	0.56	0.88	0.55	1.48
PRODUCT-A	6.5	77.11	0.43	0.74	0.55	1.69
PRODUCT-B	7.2	75.10	0.62	0.91	0.46	1.37