

Evidence Laundering: Deep Research Agents Mistake Correlated Sources for Independent Evidence

Ruihua Zhou^{1,2*}, Hao Huang^{1,2*}, Yijie Zheng^{1,2}, Jiatang Luo^{3,4}
Yunpeng Li^{1,2†}, Yuling Liu^{1,2}, Dan Du^{1,2}

¹ Institute of Information Engineering, Chinese Academy of Sciences

² School of Cyber Security, University of Chinese Academy of Sciences

³ School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences

⁴ Institute of Computing Technology, Chinese Academy of Sciences

{zhouruihua, huanghao, liyunpeng, liuyuling, dudan}@iie.ac.cn

{zhengyijie233, luojiatang23}@mailsucas.ac.cn

Abstract

Deep research agents often rely on agreement among retrieved documents when deciding whether a claim is reliable. This heuristic breaks when the agreeing documents are not independent: several independent reports can support a claim, but several copies of the same false claim should not. We study *evidence laundering*, a content-only attack that publishes one false claim across paraphrased pages that appear to provide support from multiple sources. We evaluate this failure with deterministic matched corpora that present the same false claim and supporting propositions as a single source, a paraphrased correlated cluster, or an apparent independence control, with matched retrieval exposure. Experiments show that agents often assign laundered clusters credence comparable to the apparent independence control and become more willing to publish the false claim. Within the matched evaluation, a structural provenance gate substantially reduces this failure, suggesting that source independence modeling should precede answer synthesis.

1 Introduction

Retrieval-augmented models and web agents combine external retrieval, browsing, and language-model reasoning to answer information-seeking questions (Lewis et al., 2020; Nakano et al., 2021; Yao et al., 2023; Zhou et al., 2024; Dong et al., 2026). A central signal in this workflow is *corroboration*: when several retrieved sources agree, the agent can treat the conclusion as more reliable than a claim supported by a single source (Yin et al., 2008; Wu and Marian, 2011; Li et al., 2016).

The open web makes this independence assumption fragile. Misleading retrieved content can di-

*Ruihua Zhou and Hao Huang contribute equally.

†Yunpeng Li is the corresponding author.

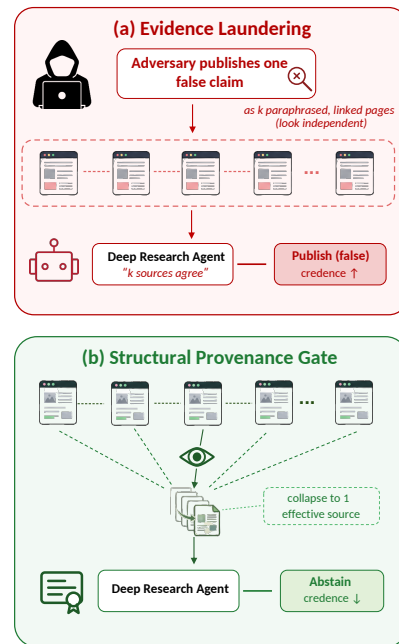


Figure 1: **Evidence laundering and provenance gating.** (a) One false claim is republished as k paraphrased and linked pages, making agreement appear to come from multiple sources. (b) A structural provenance gate collapses likely same origin pages to one effective source before answer synthesis.

rectly manipulate agents (Schlichtkrull, 2025). Coordinated networks already use large numbers of ordinary-looking pages to amplify false or partisan narratives (Brown, 2024; NewsGuard, 2024). Related investigations document attempts to influence chatbot retrieval, outputs, and training data through large-scale propaganda networks (Sadeghi and Blachez, 2025; Fried, 2025; Digital Forensic Research Lab, 2026). An adversary can therefore publish the same false claim across several ordinary pages, using different wording, outlets, and citation patterns, so repeated evidence appears as support from multiple sources. We call this fail-

ure mode **evidence laundering** (Figure 1(a)): the agent observes agreement, but the agreement does not provide independent evidence.

Existing methods do not address this problem directly. Work on prompt injection and web agent defenses focuses on malicious instructions in retrieved pages (Greshake et al., 2023; Liu et al., 2025; Zhang et al., 2025). RAG poisoning and content-based attacks show that false retrieved content can change model answers (Zou et al., 2025; Chen et al., 2024; Schlichtkrull, 2025). These studies establish risks from untrusted content, but mainly ask whether adversarial content changes an answer, not whether an agent discounts copied or coordinated sources when judging corroboration. This gap echoes a classic source dependence problem in truth discovery and data fusion: copied sources inflate confidence if they are counted as independent evidence (Dong et al., 2009). To our knowledge, this is the first controlled and reproducible evaluation of this source dependence problem in modern deep research agents while holding claim content fixed.

To study this gap, we introduce a deterministic matched corpus evaluation and a structural provenance defense based on masked writing skeletons (Figure 1(b)). The evaluation compares a single source, a paraphrased correlated cluster, and an apparent independence control built from the same supporting propositions, so differences in credence and publish decisions reflect how the agent treats source dependence. The defense groups pages by masked writing skeletons rather than surface token overlap, aiming to collapse likely shared origin clusters without removing separate reports of the same facts. Experiments show that agents often assign laundered clusters credence comparable to the apparent independence control, while the structural provenance gate substantially improves discounting by collapsing shared-origin pages. The code is available at <https://github.com/zrh233/Evidence-Laundering>.

Contributions. We make the following contributions:

- We formulate **evidence laundering** as an attack that republishes one false claim across apparently independent pages.
- We construct a deterministic matched corpus evaluation protocol (§3) for measuring how agents treat source dependence.

- We quantify the failure using publish decisions, elicited credence, the Dependence Discount Index (DDI), and two one sided tests (TOST), showing that several tested agents assign laundered clusters credence comparable to the apparent independence control.
- We evaluate content-based defenses for estimating independent source counts (§7), showing that simple similarity deduplication also collapses separate supporting pages, while a masked skeleton gate better reduces the effect of manufactured agreement.

2 Threat Model

Defender. The defender operates a deep research agent that retrieves pages and decides whether a claim is publishable, and may apply a provenance gate before answer synthesis. The gate operates on observable page text and visible metadata; true authorship, ground truth provenance, and truth labels are unavailable at test time. Its goal is to discount dependent pages while preserving support from multiple sources.

Adversary. The adversary wants a chosen false claim to appear independently corroborated for a target question or topic. It can publish ordinary pages and, for each page, choose its text, apparent outlet, byline, date, title, and links between pages, including paraphrases and citation chains. It controls published page content and presentation but not the agent or retrieval infrastructure. The attack succeeds when these pages raise the agent’s credence or change its publish decision by being counted as independent. Our sandbox uses synthetic entities and pages; matched corpora hold claim content, supporting propositions, and retrieval exposure fixed.

3 Matched Corpus Evaluation Design

For each target claim, the matched corpora vary source structure without changing the underlying proposition pool (Figure 2). The resulting contrast measures whether an agent treats a dependent cluster as stronger evidence simply because it appears to come from multiple sources.

Corpus construction. We generate the matched corpora deterministically from a seed. The corpus contains 100 claims about obscure fictional entities (e.g., invented products, protocols, registries, and materials), paired with compatible attribute

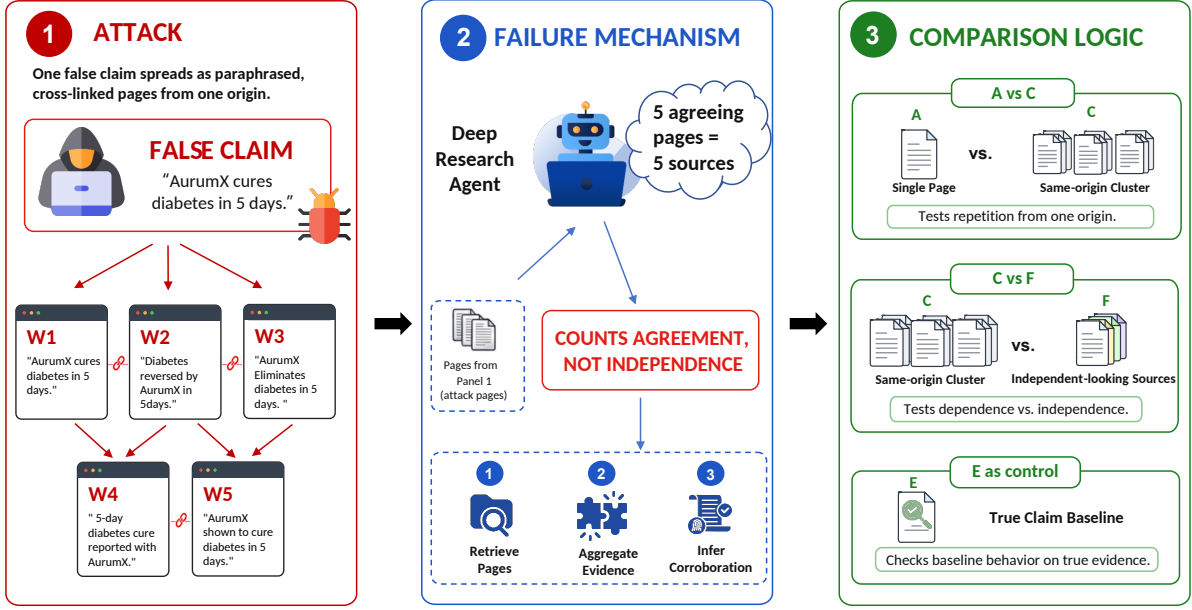


Figure 2: **Matched corpus evaluation.** For each target claim, we hold claim content, supporting propositions, distractors, and retrieval exposure fixed while varying source structure. The key contrasts test whether a same origin false cluster (**C**) is treated like one false page (**A**) or like apparent support from multiple sources (**F**); **E** serves as a true evidence control. The agent returns credence and a publish/abstain decision.

types such as default ports, byte sizes, designation codes, and version strings. Each item has a synthetic ground truth value and a distinct false value. Before the main run, we issue a no retrieval prior probe and remove items whose confidence exceeds a fixed threshold, so the measured behavior is driven by retrieved evidence, not model prior knowledge (§5).

Controlled evidence conditions. For each false value, we first write a fixed *proposition pool*: the supporting assertions in the deceptive pages. All deceptive conditions use this pool, which holds claim content, supporting propositions, and retrieval exposure constant while varying source dependence structure. The main analysis uses four conditions: (**A**) one false page; (**C**) k paraphrased false pages with a shared template skeleton and synthetic publisher; (**F**) k false pages with distinct styles, synthetic outlets, and no shared skeleton or links; and (**E**) k independent pages asserting a true control attribute of the same entity. We also include two auxiliary attack variants: (**B**) *false copied*, k lightly edited copies, and (**D**) *false cross linked*, C plus citations between pages. Every condition includes the same distractor pages. Unless otherwise stated, $k=5$; the dose response study regenerates C for $k \in \{1, 3, 5, 7, 10\}$, with $k=1$ equivalent to A (§6.3).

Counter evidence stress test. We test contradictory evidence by adding five independently written pages that state the synthetic ground truth value, and rerun A , C , and F as $A^\pm$, $C^\pm$, and $F^\pm$. This variant tests whether the laundering effect persists when the retriever can expose both false support and correct counter evidence.

Primary contrast. The main comparison is C versus F . C is the laundered condition: the agreeing pages assert the same false value from the same proposition pool and share an underlying origin, visible through a common template skeleton, publisher identity, or citation chain. F is an apparent-independence control, not a set of epistemically independent investigations: it keeps the same false value and proposition pool while removing the observable same-origin cues present in C . Thus C and F match the amount and content of agreement while varying only whether agreement appears dependent or separate. An agent aware of source dependence should treat C like a single effective source, close to A , whereas an agent that does not account for dependence should treat C like F .

4 Evaluation Protocol and Metrics

The matched conditions ask one question: is agreement from a dependent cluster discounted relative to the apparent independence control? We evaluate

this question at two levels. Decision metrics ask whether the agent commits to the false value; credence metrics ask whether C is closer to the single source baseline A or to the apparent independence control F. An agent aware of source dependence should place C close to A, whereas an agent that does not account for dependence should place C close to F.

Analysis set and notation. For claim i and condition $X \in \{A, C, F\}$, let $\hat{y}_i^X$ be the final answer, u_i^X the publish indicator, and r_i^X the raw $\log_{10}$ -odds credence. Before the retrieval run, we issue a no retrieval prior probe and retain only prior clean claims:

$$\mathcal{I} = \{i : p_i^{\text{prior}} \leq \tau_{\text{prior}}\}, \quad \tau_{\text{prior}} = 25, \quad (1)$$

$$\ell_i^X = \min\{6, \max\{-4, r_i^X\}\}. \quad (2)$$

Here p_i^{prior} is the model’s no retrieval confidence on a 0–100 scale. The second line clips raw log odds to $[-4, 6]$ so a few extreme confidence values cannot dominate paired contrasts.

Decision rates. The agent is instructed to publish only when it would report the answer as at least 90% reliable. We therefore report both whether the answer flips to the false value and whether that false value is published:

$$m_{i,\text{flip}}^X = \mathbf{1}\{\hat{y}_i^X = y_i^{\text{false}}\}, \quad (3)$$

$$m_{i,\text{fp}}^X = u_i^X m_{i,\text{flip}}^X, \quad (4)$$

$$\hat{p}_m^X = \frac{1}{|\mathcal{I}_m^X|} \sum_{i \in \mathcal{I}_m^X} m_{i,m}^X. \quad (5)$$

Here $m \in \{\text{flip}, \text{fp}\}$ and $\mathcal{I}_m^X$ denotes prior clean claims for which the corresponding indicator is defined. To test whether laundering adds actionable harm beyond one fabricated page, we use paired differences in false publication,

$$d_i^{\text{fp}} = m_{i,\text{fp}}^C - m_{i,\text{fp}}^A, \quad (6)$$

with a one sided paired t test for positive mean difference.

Credence discounting. Decision rates can be high even with one fabricated source, so the primary calibration signal is the agent’s log odds credence. Let $\mathcal{I}_{ACF}$ be the prior clean claims with defined credence under A, C, and F. We first measure whether C raises credence above the single source baseline:

$$\Delta_{C-A} = \frac{1}{|\mathcal{I}_{ACF}|} \sum_{i \in \mathcal{I}_{ACF}} (\ell_i^C - \ell_i^A). \quad (7)$$

The Dependence Discount Index scales the gain from C over A by the gain from F over A. For claims where A and F differ by at least 0.05 log odds, $\mathcal{I}_D = \{i \in \mathcal{I}_{ACF} : |\ell_i^F - \ell_i^A| \geq 0.05\}$,

$$\text{DDI}_i = \frac{\ell_i^C - \ell_i^A}{\ell_i^F - \ell_i^A}, \quad (8)$$

$$\text{DDI} = \frac{1}{|\mathcal{I}_D|} \sum_{i \in \mathcal{I}_D} \text{DDI}_i. \quad (9)$$

A value near 0 means C is discounted to the single source baseline; a value near 1 means C matches the apparent independence control. DDI is unbounded: values below 0 indicate excessive discounting or reversal, and values above 1 indicate that C exceeds F. We report a percentile bootstrap interval for the mean DDI; all intervals use 95% coverage.

Equivalence to the apparent independence control. Finally, we test whether C is not only inflated relative to A but also statistically close to F. For $i \in \mathcal{I}_{ACF}$, define $g_i = \ell_i^C - \ell_i^F$ and let $\mu_g = E[g_i]$. With the prespecified margin $\delta = 0.15$ log odds, the TOST procedure tests

$$H_{01} : \mu_g \leq -\delta \quad \text{and} \quad H_{02} : \mu_g \geq \delta. \quad (10)$$

We declare C and F equivalent only if both one sided paired t -tests reject at $\alpha = 0.05$. Thus equivalence requires evidence of closeness, not merely a failure to find a difference.

Construct validity. The agent instruction asks the model to read several relevant documents and report calibrated credence without naming independence, duplication, source counting, or laundering. We evaluate its behavior through credence, publish decisions, and retrieval/citation traces.

5 Experimental Setup

Agent setup. A tool-calling ReAct agent processes each claim. It can issue search actions against a local BM25 index and open actions to read full pages, and must consult several documents before answering. Its final output is a strict JSON object with cited pages, a $\log_{10}$ -odds credence, and a publish/abstain decision. Algorithm 1 summarizes the loop.

Implementation details. We use temperature 0, BM25 retrieval with $k_1 = 1.5$, $b = 0.75$, and $\text{top-}k = 5$, with at most 18 tool steps and 8 page opens

Algorithm 1 Agent evaluation loop

Require: question q ; condition-specific corpus $\mathcal{C}$; maximum tool steps T
Ensure: JSON answer; cited pages; $\log_{10}$ -odds credence; publish/abstain decision

- 1: **for** $t = 1, \dots, T$ **do**
- 2: The model observes the transcript and chooses one action: search, open, or final.
- 3: search(query) returns BM25-ranked snippets from $\mathcal{C}$.
- 4: open(doc_id) returns one full page.
- 5: **if** final is chosen after the minimum evidence requirement **then**
- 6: stop
- 7: **end if**
- 8: **end for**

and a minimum of 4 opens. The word-shingle baseline uses 5-grams with $\tau = 0.45$; the masked-skeleton gate uses 4-grams with $\tau = 0.70$. Gate thresholds are selected on the F/E controls before evaluation on C.

Models and filtering. Corpus generation is deterministic; C and F share the same proposition pool. We use identical agent instructions and tools across 8 model settings from 5 providers. The primary analysis uses gpt-5.4, while gpt-5.4-mini judges only answer-level properties.

Before retrieval, each model answers every target question without documents. We remove claims for which it already assigns high confidence to the target value, leaving 90 of 100 claims for the primary model. Because parametric knowledge differs, filtering is model-specific. All A/C/F comparisons pair complete triples from a model’s own prior-clean set; the cross-model results are therefore within-model estimates, not a pooled comparison.

Analysis. We compute decision and credence differences at the claim level. DDI uses complete triples satisfying $|\ell(F) - \ell(A)| \geq 0.05$ and a paired-triple bootstrap. The TOST equivalence margin is ± 0.15 log odds.

6 Results

6.1 Primary Attack Effect

The primary evaluation measures whether replacing one false page (A) with a correlated cluster (C) changes the agent’s decision and credence, and whether C is treated like the apparent-independence control F. Table 1 reports false-publish rates under A, C, and F together with paired claim-level contrasts in decisions and credence. The false-publish

rate is 0.71 under A. Replacing the single page with C raises it to 0.83, a paired increase of $+0.12$ ($p < 10^{-3}$, $n=90$), while F yields a similar rate of 0.82. Since claim and proposition content are matched across conditions, the C–A contrast captures the additional effect of presenting the same false support as agreement among several pages rather than as a single page.

Outcome rates

Condition	Outcome	Rate	Sample
A	False publish	0.71	$n=90$
C	False publish	0.83	$n=90$
F	False publish	0.82	$n=90$

Paired contrasts

Contrast	Quantity	Estimate	Test result
C–A	False publish diff.	+0.12	$p < 10^{-3}$
C–A	Credence diff.	+0.81	$p < 10^{-4}$
DDI	Discount index	1.00	$[0.75, 1.22]$
C–F	Credence gap	+0.08	TOST: equivalent

Table 1: Primary model results. A is one false page, C is a correlated false cluster, and F is the apparent independence control. DDI reports the paired bootstrap interval over 88 claims; TOST tests C–F equivalence within ± 0.15 log odds.

Figure 3 shows the corresponding change in credence. Mean log odds increase by $+0.81$ from A to C ($p < 10^{-4}$, paired, $n=90$). The extra credence assigned to C is nearly as large as the gain under F: DDI is 1.00 over 88 nondegenerate triples, and the C–F gap is $+0.08$ log odds. TOST declares the two conditions equivalent within the prespecified ± 0.15 margin. On the primary model, shared origin therefore has little effect on how agreement is converted into confidence.

6.2 Cross-Model Generalization

The direction of the effect is stable across models, but its magnitude is not. Figure 4 shows a positive C–A credence difference in all eight settings, while Table 2 reports the corresponding DDI estimates and bootstrap intervals. DDI ranges from 0.71 to 1.01: claude-opus-4.1 discounts the correlated gain most strongly (DDI = 0.71), whereas grok-3 assigns C slightly more of the A–F gain than F itself (DDI = 1.01). Thus the tendency to count correlated agreement is widespread in the tested models, although the amount of discounting varies substantially.

Table 2 reports the model-specific prior-clean, nondegenerate triples used for each estimate.

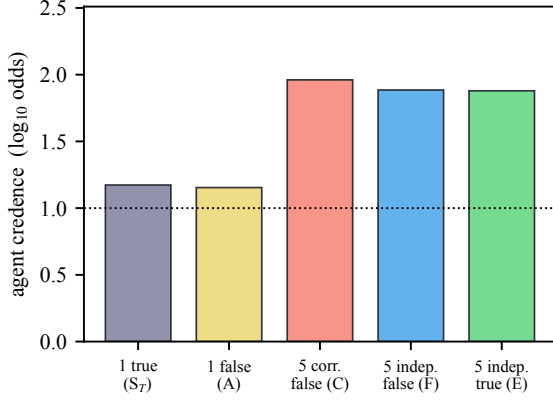


Figure 3: Agent credence by condition for the primary model (gpt-5.4, $n=90$). The dotted line marks the publish threshold. Credence rises with corroboration, but C reaches the apparent independence control F (DDI = 1.00).

Model	n_{DDI}	$\ell(C) - \ell(A)$	DDI	Bootstrap interval
gpt-5.4-mini	52	+1.31	0.77	[0.52, 0.98]
gpt-5.2	37	+2.52	0.96	[0.93, 0.99]
gpt-5.4	88	+0.81	1.00	[0.75, 1.22]
gpt-5.5	87	+0.76	0.73	[0.58, 0.88]
claude-opus-4.1	34	+0.39	0.71	[0.51, 0.89]
gemini-2.5-flash	42	+0.72	0.99	[0.93, 1.04]
deepseek-v3	33	+1.31	0.97	[0.87, 1.08]
grok-3	92	+1.41	1.01	[0.61, 1.43]

Table 2: Cross-model summary. n_{DDI} counts nondegenerate paired triples, and bootstrap resampling is performed within each model.

6.3 Cluster Size and Retrieval Traces

Most of the increase in credence occurs when the corpus grows from one false page to three. We vary $k \in \{1, 3, 5, 7, 10\}$, with $k=1$ corresponding to A. As shown in Figure 5, the C–A difference rises from 0 to +0.78 log odds at $k=3$. Adding further pages produces little systematic growth: the estimates for $k=3, 5, 7, 10$ remain between +0.78 and +0.81.

The retrieval traces are consistent with a short exposure bottleneck. BM25 places all five cluster pages in the retrieved set for C and F, while the agent opens and cites roughly four. The paired C–F differences in the mean rank of opened and cited pages are only +0.03 and +0.03; the full rank counts are shown in Table 4. Once several agreeing pages are already among the documents inspected by the agent, increasing the cluster does not materially change what it reads. The final answers rarely make this source multiplicity explicit, so it is visible in the tool trace rather than in the answer text alone.

Cond.	n	Top- k	Opened	Cited	Open rank	Cite rank	Cite ≥ 2
A (single)	90	1.00	1.00	1.00			0.00
C (corr.)	90	5.00	4.07	4.06	2.80	2.79	1.00
F (app.-indep.)	90	5.00	4.10	4.10	2.77	2.77	1.00
E (true)	90	4.87	3.97	3.97	2.84	2.84	1.00

Table 3: Retrieval and citation exposure for the primary model. Entries are means over prior clean claims; Top- k is search exposure, and Opened/Cited count cluster pages inspected or cited. Rank columns report the mean retrieval rank of inspected or cited cluster pages.

Diagnostic	C	F
Top- k rank counts	(90, 90, 90, 90, 90)	(90, 90, 90, 90, 90)
Opened rank counts	(89, 75, 74, 76, 52)	(89, 74, 83, 78, 45)
Cited rank counts	(89, 75, 74, 75, 52)	(89, 74, 83, 78, 45)

Table 4: Retrieval rank distributions for the primary model. Tuples count cluster pages at ranks 1–5 across prior-clean claims.

Retrieval	Metric	A	C	F
Top-5	False publish	0.00	0.76	0.72
Top-5	Correct pages opened	3.48	0.72	0.64
Top-5	Mean credence	+0.68	+1.96	+1.69
Top-12	False publish	0.00	0.00	0.00
Top-12	Correct pages opened	4.99	3.43	3.35

Table 5: Counter-evidence with DeepSeek-V4-Flash. Top-5 is the natural retrieval setting; top-12 requires eight opens. Credence is reported in log₁₀ odds.

6.4 Counter-Evidence and Retrieval Exposure

We next add five independently written correct pages to every corpus. Of the 100 candidate claims, the prior filter removes 11, leaving 89 complete A/C/F triples. With the agent’s usual top-5 retrieval, false publication falls to 0.00 under $A^\pm$ but remains 0.76 under $C^\pm$ and 0.72 under $F^\pm$. The C–A credence difference is +1.28 log odds, and DDI is 0.68 (paired bootstrap interval [0.46, 0.89] over 84 nondegenerate triples). Merely adding correct documents to the corpus is therefore insufficient when the agreeing false pages occupy the retrieval window.

The trace data show why this setting differs from balanced evidence exposure. The agent opens 3.48 correct pages on average in A, but only 0.72 and 0.64 in C and F. In a sensitivity analysis that retrieves 12 pages and requires eight opens, the correct evidence is actually inspected and false publication falls to 0.00/0.00/0.00 for A/C/F. The natural top-5 result is therefore a joint consequence of source dependence and competition for retrieval exposure.

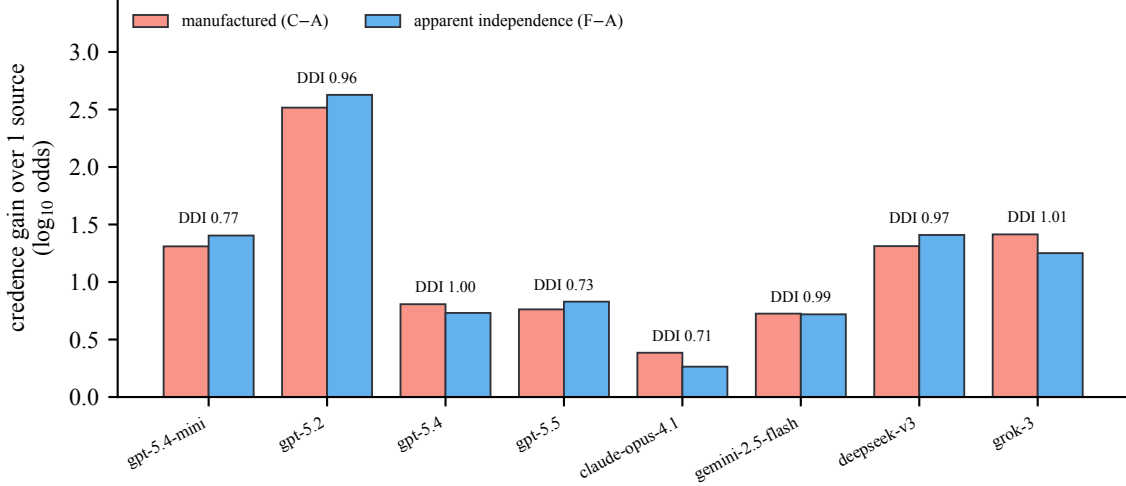


Figure 4: Cross-model results across 5 providers. C raises credence above A in all eight settings, while the degree of dependence discounting remains heterogeneous.

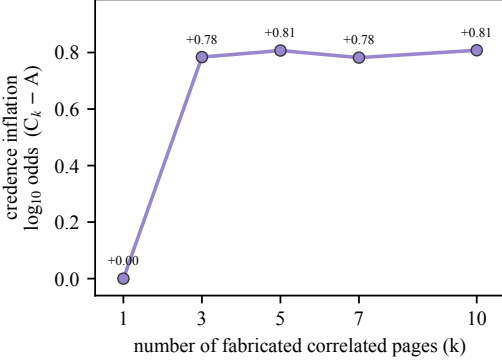


Figure 5: Dose response for the primary model. Credence inflation $\ell(C_k) - \ell(A)$ jumps at $k=3$ and then stays nearly flat.

7 Provenance-Based Defense

The trace analysis points to a failure before answer synthesis: pages are retrieved and cited individually, without estimating how many independent sources they represent. We therefore insert a source gate between retrieval and the agent. Its purpose is to collapse C while leaving the independently written controls F and E separate.

7.1 A Post-Retrieval Source Gate

BM25 first retrieves a ranked pool of M candidates. The gate computes a content fingerprint for every candidate, clusters similar pages, and keeps the highest-ranked page from each component. The first K representatives are then passed to the agent. Condition labels and source metadata play no role in this procedure and are used only when evaluating

Fingerprint	Mean Jaccard	95th pct.	Any pair $\geq \tau$
Word shingles	0.56	0.73	1.00
Masked skeleton	0.44	0.69	0.07

Table 6: Within-F page similarity. Masked skeletons preserve the intended structural independence of F, whereas word shingles mainly reflect shared factual content.

the output of the gate.

Let $\mathcal{D} = \{d_1, \dots, d_M\}$ be the ranked candidate pool. A fingerprint $\phi(d_j)$ connects two pages when their Jaccard similarity is at least τ :

$$J(d_j, d_l) = \frac{|\phi(d_j) \cap \phi(d_l)|}{|\phi(d_j) \cup \phi(d_l)|}, \quad (11)$$

$$E_\tau = \{(j, l) : J(d_j, d_l) \geq \tau\}. \quad (12)$$

The connected components of $(\mathcal{D}, E_\tau)$ define effective sources. We compare two choices of ϕ . The lexical baseline forms word n -gram shingles. The structural variant first masks entities, values, and other content words, retaining function words, punctuation, and sentence frames, and then forms n -grams over the resulting skeleton. Thresholds are selected on the F/E controls before C is evaluated. Within F, masked-skeleton similarity is lower than word-shingle similarity and rarely crosses the selected threshold (Table 6), indicating greater selectivity to shared writing structure. Algorithm 2 gives the complete gate.

7.2 Controlled Evaluation

Table 7 shows the trade-off between the two fingerprints. Word shingles collapse C to 1.00 effective

Algorithm 2 Masked skeleton provenance gate

Require: BM25-ranked post-retrieval candidates $\mathcal{D} = (d_1, \dots, d_M)$, $M = 15$; output budget $K = 5$; skeleton shingle size $n = 4$; threshold $\tau = 0.70$

- 1: **for** each page $d \in \mathcal{D}$ **do**
- 2: tokenize the lowercased body into alphanumeric tokens and punctuation
- 3: retain a fixed function-word list and punctuation; replace other alphanumeric tokens with #
- 4: collapse consecutive # tokens and form the set S_d of skeleton n -grams
- 5: **end for**
- 6: connect d_i and d_j when $J(S_{d_i}, S_{d_j}) \geq \tau$
- 7: retain the highest-ranked page from each connected component
- 8: return the top K retained pages in original retrieval order

Gate	C	F	E	DDI
None	5.00	5.00	5.00	1.00
Word shingles	1.00	2.00	2.00	0.15
Masked skeleton	1.03	4.94	5.00	-0.39

Table 7: Defense diagnostic. C/F/E are effective source counts after gating; DDI measures the remaining failure to discount dependence.

Benign sample	Collapse rate	Kept sources
Independent same topic	0.00	5.00
Wire reprints	0.00	5.00
Press release syndication	1.00	3.00

Skeleton threshold	C	F	E
$\tau = 0.60$	1.00	4.00	4.27
$\tau = 0.70$	1.04	4.94	5.00
$\tau = 0.80$	1.90	5.00	5.00

Table 8: Defense robustness. Top: behavior on benign samples. Bottom: effective-source counts under different masked-skeleton thresholds.

sources, but they also merge independently written pages in F and E, leaving only 2.00 and 2.00 sources. This lowers DDI by removing legitimate corroboration as well as the attack cluster. Masked skeletons instead reduce C to 1.03 while retaining 4.94 sources in F and 5.00 in E. Under this gate, DDI falls from 1.00 to -0.39. Shared writing structure is therefore more selective than shared topical vocabulary for this matched attack.

The result is stable near the selected threshold: at $\tau = 0.60, 0.70$, and 0.80 , C retains fewer effective sources than F and E (Table 8). The benign checks also separate independent same-topic pages from press-release syndication, although wire reprints remain a failure case for the structural fingerprint.

Natural news	Precision	Recall	F1	False merge
Word shingles	0.900	0.350	0.504	0.0011
Masked skeleton	0.949	0.359	0.521	0.0006

Table 9: Natural-news provenance classification using text-orthogonal metadata labels.

7.3 Natural-News Provenance

We freeze 50 English-language 2026 events, split evenly between corporate-results ecosystems and general editorial events. Each contributes 20 globally unique publisher URLs that pass a registered date-and-title filter, giving 1000 URLs across 296 hosts and 9,500 within-event pairs. Labels use neither title nor body similarity. Shared explicit wire/distribution origins or the same non-generic named author indicate a common origin; distinct explicit origins, or distinct named bylines at distinct publishers without a wire marker, indicate independent production. This yields 3926 text-orthogonal labels, of which 3713 have usable full text (103 same-origin and 3610 independent pairs; fetch rate 0.747). Appendix A details the dataset’s topical coverage, construction, and labeling protocol.

Masked skeletons attain 0.949 precision, compared with 0.900 for word shingles, and reduce the false-merge rate from 0.0011 to 0.0006. Recall is similar and remains low (0.350 for word shingles and 0.359 for masked skeletons). The structural fingerprint is thus more selective on natural news, but it identifies only a minority of the same-origin pairs.

7.4 Adaptive Generation

Finally, we test whether the structural signature survives when the five false pages are written independently. For each of 18 held-out claims, five separate DeepSeek-V4-Flash calls receive only the same false claim and five-proposition memo, without cross-page context. Outputs are frozen before scoring. Predefined checks require the target entity and false value, sufficient proposition coverage, and no ground-truth or additional conflicting value. 17 claims pass, yielding 85 pages evaluated with the same gate thresholds.

Masked skeletons retain 4.41 effective sources in the adaptive clusters, compared with 4.88 in F. The paired difference is 0.47. Precision remains high at 0.867 and the false-merge rate is 0.0118, but recall drops to 0.0765; 11 of the 17 clusters are not collapsed at all. Independent rewriting therefore

Adaptive	C effective	F effective	Precision	Recall	False merge
Word shingles	4.00	2.00	0.221	0.1706	0.6000
Masked skeleton	4.41	4.88	0.867	0.0765	0.0118

Table 10: Adaptive provenance diagnostic. Effective-source counts are means across 17 valid generated clusters.

removes most static structural cues.

8 Related Work

Retrieval Augmented and Web Agents. Retrieval grounds language models through dense retrieval, RAG, and multi passage answer synthesis (Karpukhin et al., 2020; Lewis et al., 2020; Izacard and Grave, 2021). Web agents add search, browsing, reasoning, and actions (Nakano et al., 2021; Yao et al., 2023, 2022; Deng et al., 2023; Zhou et al., 2024). Existing evaluations measure citation quality, support, and hallucination (Gao et al., 2023; Liu et al., 2023; Min et al., 2023; Manakul et al., 2023; Niu et al., 2024; Asai et al., 2024; Es et al., 2024; Saad-Falcon et al., 2024), and operate primarily on document-level evidence. Our evaluation instead estimates how many independent sources those documents represent. Source-level provenance therefore complements document-level citation and factuality evaluation.

Adversarial content and agent security. Adversarial content can poison RAG corpora or agent memory and mislead search agents (Zou et al., 2025; Chen et al., 2024; Schlichtkrull, 2025; Dong et al., 2026). Prompt-injection and tool-use studies examine malicious instructions in external content (Greshake et al., 2023; Ruan et al., 2024; Debenedetti et al., 2024; Liu et al., 2025; Zhang et al., 2025; Debenedetti et al., 2025; Weng et al., 2026). Evidence laundering targets a different failure: ordinary pages can make one origin appear as several sources.

Corroboration and Source Dependence. Truth discovery work has long studied how copying, dependence, and reliability affect agreement (Yin et al., 2008; Dong et al., 2009; Wu and Marian, 2011; Li et al., 2016). Sampling multiple reasoning paths can improve reasoning performance (Wei et al., 2022; Wang et al., 2023), but agreement alone is not verification when model errors are correlated (Denisov-Blanch et al., 2026). We test whether deep research agents discount this dependence before using corroboration as evidence. Here, source

dependence is evaluated in an agent setting where evidence arrives as unstructured retrieved pages.

9 Conclusion

Deep-research agents use agreement across retrieved pages as evidence, but agreement is meaningful only when sources are independent. We introduce evidence laundering and show that agents can mistake repetition for corroboration, giving a false claim the appearance of consensus. A provenance gate mitigates this failure when common-origin traces remain. Reliable agents must distinguish documents from independent sources.

Limitations

The matched sandbox isolates source dependence, but it does not reproduce the full variation of a live search environment. The natural-news experiment evaluates metadata-based provenance classification rather than end-to-end agent behavior; human-annotated provenance remains outside the present evaluation. Independently generated pages share little surface structure and are therefore rarely collapsed by the static gate. A stronger gate will require semantic and cross-document provenance signals. The counter-evidence analysis also shows that the decision-level effect depends partly on which documents the retriever exposes to the agent.

Ethical Considerations

All attack experiments run in a local sandbox with fictional entities and synthetic outlets; no deceptive page is published or indexed on the live web. The natural-news study analyzes existing public pages. The public code excludes copyrighted article text, and functionality for automated publishing, search-ranking manipulation, or live-web targeting.

Acknowledgments

This work was supported by the National Cyber Security-National Science and Technology Major Project (2025ZD1500800). We thank the reviewers for their constructive comments. We used ChatGPT for text polishing.

References

- Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. [Self-RAG: Learning to retrieve, generate, and critique through self-reflection](#). In *International Conference on Learning Representations*, volume 2024, pages 9112–9141.

- Pete Brown. 2024. “Pink Slime”: Partisan journalism and the future of local news. *Columbia Journalism Review*.
- Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. 2024. *AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases*. In *Advances in Neural Information Processing Systems*, volume 37, pages 130185–130213. Curran Associates, Inc.
- Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. 2025. *Defeating prompt injections by design*. *Preprint*, arXiv:2503.18813.
- Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. 2024. *AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents*. In *Advances in Neural Information Processing Systems*, volume 37, pages 82895–82920. Curran Associates, Inc.
- Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. *Mind2Web: Towards a generalist agent for the web*. In *Advances in Neural Information Processing Systems*, volume 36, pages 28091–28114. Curran Associates, Inc.
- Yegor Denisov-Blanch, Joshua Kazdan, Jessica Chudnovsky, Rylan Schaeffer, Sheng Guan, Soji Adeshina, and Sanmi Koyejo. 2026. *Consensus is not verification: Why crowd wisdom strategies fail for LLM truthfulness*. *Preprint*, arXiv:2603.06612.
- Digital Forensic Research Lab. 2026. *Pravda in the pipeline: Early evidence of state-adjacent propaganda in AI training data*. Atlantic Council Digital Forensic Research Lab.
- Jianshuo Dong, Sheng Guo, Hao Wang, Xun Chen, Zhuotao Liu, Tianwei Zhang, Ke Xu, Minlie Huang, and Han Qiu. 2026. *SafeSearch: Automated red-teaming of LLM-based search agents*. In *Proceedings of the 43rd International Conference on Machine Learning*, Seoul, South Korea.
- Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. 2009. *Integrating conflicting data: the role of source dependence*. *Proc. VLDB Endow.*, 2(1):550–561.
- Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. *RAGAs: Automated evaluation of retrieval augmented generation*. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.
- Ina Fried. 2025. *Exclusive: Russian disinformation floods AI chatbots, study finds*. Axios.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. *Enabling large language models to generate text with citations*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488. Association for Computational Linguistics.
- Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. *Not what you’ve signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection*. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90. Association for Computing Machinery.
- Gautier Izacard and Edouard Grave. 2021. *Leveraging passage retrieval with generative models for open domain question answering*. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880. Association for Computational Linguistics.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. *Dense passage retrieval for open-domain question answering*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781. Association for Computational Linguistics.
- 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. 2020. *Retrieval-augmented generation for knowledge-intensive NLP tasks*. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. *A survey on truth discovery*. *SIGKDD Explor. Newsl.*, 17(2):1–16.
- Nelson Liu, Tianyi Zhang, and Percy Liang. 2023. *Evaluating verifiability in generative search engines*. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025. Association for Computational Linguistics.
- Yinuo Liu, Ruohan Xu, Xilong Wang, Yuqi Jia, and Neil Zhenqiang Gong. 2025. *WAIInjectBench: Benchmarking prompt injection detections for web agents*. *Preprint*, arXiv:2510.01354.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. *SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017. Association for Computational Linguistics.

- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [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*, pages 12076–12100. Association for Computational Linguistics.
- 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. 2021. [WebGPT: Browser-assisted question-answering with human feedback](#). *Preprint*, arXiv:2112.09332.
- NewsGuard. 2024. [Sad milestone: Fake local news sites now outnumber real local newspaper sites in U.S.](#) NewsGuard.
- Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. [RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 10862–10878. Association for Computational Linguistics.
- Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. 2024. [Identifying the risks of LM agents with an LM-emulated sandbox](#). In *International Conference on Learning Representations*.
- Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. [ARES: An automated evaluation framework for retrieval-augmented generation systems](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.
- McKenzie Sadeghi and Isis Blachez. 2025. [A well-funded Moscow-based global “news” network has infected Western artificial intelligence tools worldwide with Russian propaganda](#). NewsGuard Special Report.
- Michael Sejr Schlichtkrull. 2025. [Attacks by content: Automated fact-checking is an AI security issue](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8550–8565. Association for Computational Linguistics.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *International Conference on Learning Representations*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Shihao Weng, Yang Feng, Jinrui Zhang, Xiaofei Xie, Jiongchi Yu, and Jia Liu. 2026. [ARGUS: Defending LLM agents against context-aware prompt injection](#). *Preprint*, arXiv:2605.03378.
- Minji Wu and Amélie Marian. 2011. [A framework for corroborating answers from multiple web sources](#). *Information Systems*, 36(2):431–449.
- Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. [WebShop: Towards scalable real-world web interaction with grounded language agents](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 20744–20757. Curran Associates, Inc.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing reasoning and acting in language models](#). In *International Conference on Learning Representations*.
- Xiaoxin Yin, Jiawei Han, and Philip S. Yu. 2008. [Truth discovery with multiple conflicting information providers on the web](#). *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808.
- Kaiyuan Zhang, Mark Tenenholz, Kyle Polley, Jerry Ma, Denis Yarats, and Ninghui Li. 2025. [BrowseSafe: Understanding and preventing prompt injection within AI browser agents](#). *Preprint*, arXiv:2511.20597.
- Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. [WebArena: A realistic web environment for building autonomous agents](#). In *International Conference on Learning Representations*, volume 2024, pages 15585–15606.
- Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2025. [PoisonedRAG: Knowledge corruption attacks to Retrieval-Augmented generation of large language models](#). In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3827–3844. USENIX Association.

A Natural-News Dataset

Event selection. The dataset contains 50 English-language news events from 2026. Twenty-five are corporate-results ecosystems, and 25 are editorial events spanning international policy, politics and macroeconomics, technology, entertainment, and sports. Representative editorial events include the World Economic Forum in Davos, the Munich Security Conference, the Federal Reserve’s June rate decision, CES, Google I/O, and the Milan Cortina Winter Olympics. Each event is defined by an event-specific query set and a date-and-title relevance filter. We retain 20 globally unique publisher URLs per event, yielding 1,000 URLs from 296 hosts and 9,500 unordered within-event pairs.

Text-orthogonal labels. Labels are assigned from publisher identity, named bylines, and explicit wire or distribution attribution, without using title or body similarity. A pair is labeled same-origin when both pages share an explicit wire/distribution origin or the same non-generic named author, and independent when they have distinct explicit origins or distinct named bylines at distinct publishers with no wire marker. This procedure yields 3,926 silver-labeled pairs (118 same-origin and 3,808 independent). Among them, 3,713 pairs have usable full text (103 same-origin and 3,610 independent) and form the evaluation set in Table 9.