RECORD GROUPING CONTROLS EVIDENCE WEIGHT IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeated retrieval records can make one evidence source appear to provide several independent contributions. We group records before generation, remove exact within-group copies, and preserve complementary content in one record per supplied group. In the central natural-text experiment, replacing four falsely split records with one full-content group reduces the normalized likelihood share of the attacker-favored candidate by 18.2–40.7% across all eight model–dataset pairs, while retaining the same 160 evidence words. A stricter six-slot control holds headers, object count, and token length fixed and yields reductions of 1.2–15.6%. On 120 real questions with overlapping distractor excerpts, estimated document grouping improves answer accuracy by 5.42 and 5.83 percentage points for Mistral-7B and Phi-4.

1 INTRODUCTION

Retrieval-augmented systems consume records whose multiplicity can diverge from the number of underlying evidence sources. A coordinated campaign can render one information-generating process across many pages, exposing each rendering as a separate contribution opportunity. Chunking, syndication handling, and record emission therefore influence how evidence reaches the model.

Prior work has established that frequency, repeated arguments, source labels, metadata, and document diversity can steer language-model decisions (Jin et al., 2024; Wan et al., 2024; Chiang & Lee, 2024; Schuster et al., 2026; Naphade, 2026; Ross et al., 2026). We take this behavioral susceptibility as the starting point. Once an upstream partition specifies which records share a dependence unit, a group–content representation can preserve complementary content while removing within-group copies. Exact copies then leave that representation unchanged. A bounded external aggregator additionally limits each group's numerical contribution; the response of a language model to serialized group records is measured empirically.

Group count alone leaves an essential choice unresolved. Four distinct evidence elements can be grouped as $\{\{1,2\},\{3,4\}\}$ or $\{\{1,3\},\{2,4\}\}$: both have two groups of size two, yet combine different facts within each unit. We prove that even bounded group aggregation can distinguish these states. This motivates retaining group membership together with content.

We group records before generation. An upstream process supplies the partition. Model-facing membership labels are opaque: equality encodes membership, while authenticated identity or credibility metadata, when available, live in explicit group signatures. Our text implementation removes exact copies, concatenates complementary passages, and emits one record per group. Figure 1 shows the measured reduction in attacker-answer probability alongside this intervention.

The central comparison keeps all 160 evidence words and changes their grouping, reducing attacker-answer probability by 18.2–40.7% across eight model–dataset pairs (Table 1). This probability measures preference between two specified candidates. A stricter six-slot control fixes repeated fields and tokenizer length and yields 1.2–15.6% reductions. Complementary-content and independent-root controls separate grouping from content removal; campaign renderings reveal model and order reversals.

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.

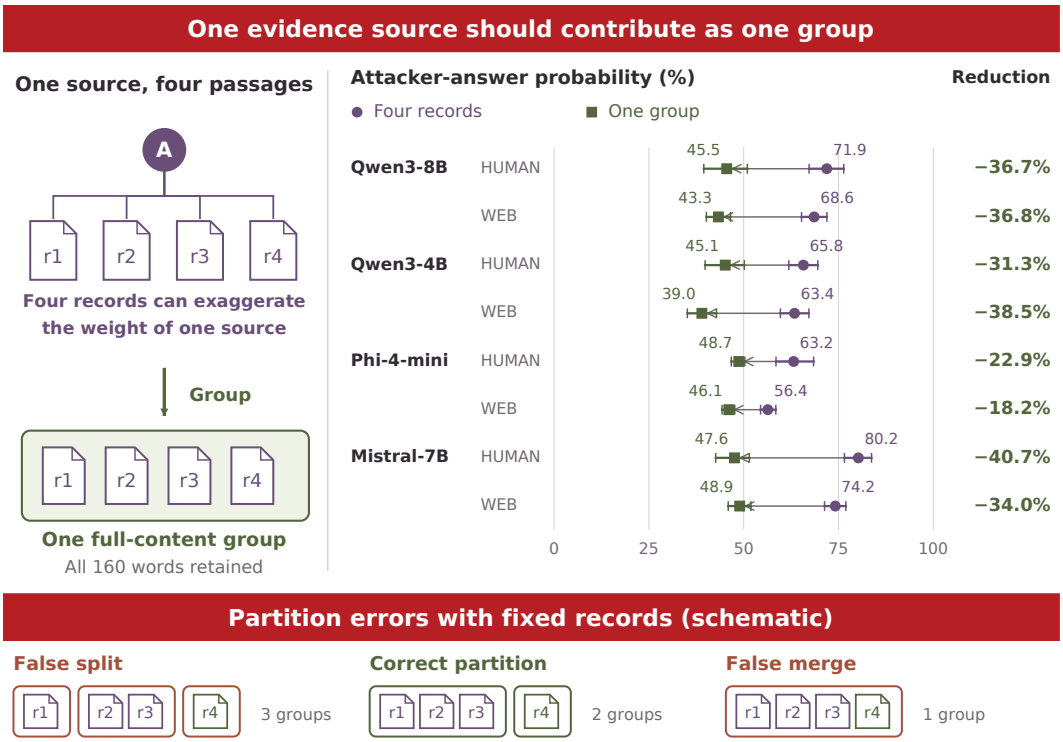


Figure 1: Why group records before generation? Four retrieved passages can originate from one evidence source, yet separate records give that source repeated opportunities to influence the model. The left diagram groups four 40-word passages while retaining all 160 words. In the plot, attacker-answer probability is the normalized likelihood assigned to the answer favored by the manipulated evidence; lower values mean less preference for that answer. Circles show four separate records and squares show one full-content group for each model on the HUMAN and WEB datasets. Error bars are pointwise 95% dependence-component bootstrap intervals; relative reductions are 18.2–40.7% (Table 1). The lower schematics use purple records from one source and green from another: a false split counts one source in multiple groups, while a false merge combines distinct sources.

Our contributions are threefold.

- We give a pre-generation group–content representation that retains complementary content and removes exact within-group copies. Authenticated identity remains available through explicit signatures; the text implementation emits one record per supplied group.
- We characterize invariant and faithful group–content states and derive a sharp content-aware error bound. The same four records arranged into two groups can yield different bounded evidence states, showing why group count alone is insufficient.
- We measure an 18.2–40.7% reduction in attacker-answer probability when four same-root records become one full-content group, with all 160 evidence words retained. A stricter matched six-slot control retains 1.2–15.6% reductions. Complementary-content, independent-root, and cross-genre campaign comparisons identify what grouping preserves and when model responses reverse.

2 Related Work

Context conflicts and source preferences. Language models arbitrate inconsistently between parametric memory, contextual claims, and user-supplied specifications (Longpre et al., 2021; Zhou et al., 2023; 2024). Evidence frequency, rationale framing, metadata, institutional labels, and source authority further alter this balance (Jin et al., 2024; Sun et al., 2026; Wan et al., 2024; Chiang & Lee,

2024; Ge et al., 2025; Li et al., 2025; Liao, 2026). These studies establish that contextual frequency and source cues are behaviorally active. We enforce replication invariance in the model-facing representation before generation.

Redundancy, diversity, and grouped evidence. GroupQA shows that paraphrased documents supporting one argument can outweigh distinct support and documents order effects and unfaithful explanations (Naphade, 2026). Whose Facts Win? shows that repetition can reverse credibility preferences across 13 open-weight models and combines teacher–student LoRA distillation with a credibility-aware prompt to induce approximate repetition invariance (Schuster et al., 2026). In a benign fictional-QA setting, Ross et al. (2026) find little correctness gain from duplicates or paraphrases and a large gain from diverse documents. Song et al. (2026) attach canonical identities to observations and maintain mergeable aggregation states outside long-context generation. We study the representation itself: supplied source-dependence keys induce a group–content state before generation, and content-fixed split/merge interventions with matched six-slot serialization identify the effect of record partition on evidence weight.

Adversarial retrieval and source grouping. SearchGEO studies adversarial web content that corrupts search-agent endorsements (Chen et al., 2026); generative-engine optimization studies how content changes increase visibility in generated answers (Aggarwal et al., 2024). Rahadi (2026) proposes provenance-graph estimation, effective independent evidence count and confidence-inflation diagnostics, and copy-cluster-discounted answer aggregation. Its aggregation uses membership information as well as counts. We analyze pre-generation invariance, within-group content retention, and distortion from partition error. Equal-count separation below concerns scalar count summaries; the controlled interventions establish checkpoint-dependent responses to record boundaries. Adversarial tool outputs expose the same problem for agents (Zhan et al., 2026), while instruction-hierarchy training improves prioritization of privileged over untrusted inputs (Wallace et al., 2024).

3 Partitioned Evidence Before Generation

3.1 Claim-Relative Evidence Roots

For a target claim c and fixed observed record set I, let $\rho_c(r) \in \mathcal{S}_c$ denote the information-generating root from which record r derives its claim-relevant evidence. Each $r \in I$ is a claim-evidence atom with exactly one claim-relative root. A physical document drawing on multiple roots enters I only after atomization into such units or assignment to an explicitly defined composite root. Here "independent" denotes distinct roots under this provenance relation; statistical dependence among their observations remains admissible. The corresponding oracle partition is

$$P_c^\star = \big\{\{r \in I : \rho_c(r) = s\} : s \in \rho_c(I)\big\}. \tag{1}$$

When admissible worlds vary below, $P_c^\star(w)$ denotes the partition obtained in world w; elsewhere we suppress w. Its block count is the *epistemic multiplicity* of the retrieved evidence for c; the number of returned records is its presentation multiplicity. The distinction is claim-relative: two articles can share a measurement for one claim while carrying independently produced evidence for another.

An upstream dependence-discovery process estimates $P_c^\star$ as $\widehat{P}_c$. The downstream mechanism accepts either partition as an input. Opaque labels encode membership; authenticated signatures carry provenance and credibility. Under $P_c^\star$, another rendering from an existing root may enrich canonical content inside that block while the group-level summand count stays fixed.

For a fixed claim and record count, let $Z_c(w)$ be the complete visible text in an admissible world w. Universal deterministic text-only recovery exists exactly when $Z_c(w_0) = Z_c(w_1)$ implies $P_c^\star(w_0) = P_c^\star(w_1)$ (Proposition 6).

Equivalently, the oracle partition is constant on every observational fiber of Z_c. Similarity models provide useful evidence under distributional assumptions; universal exact recovery is available precisely when worlds with identical visible text share the same oracle partition. Appendix B.1 proves the characterization.

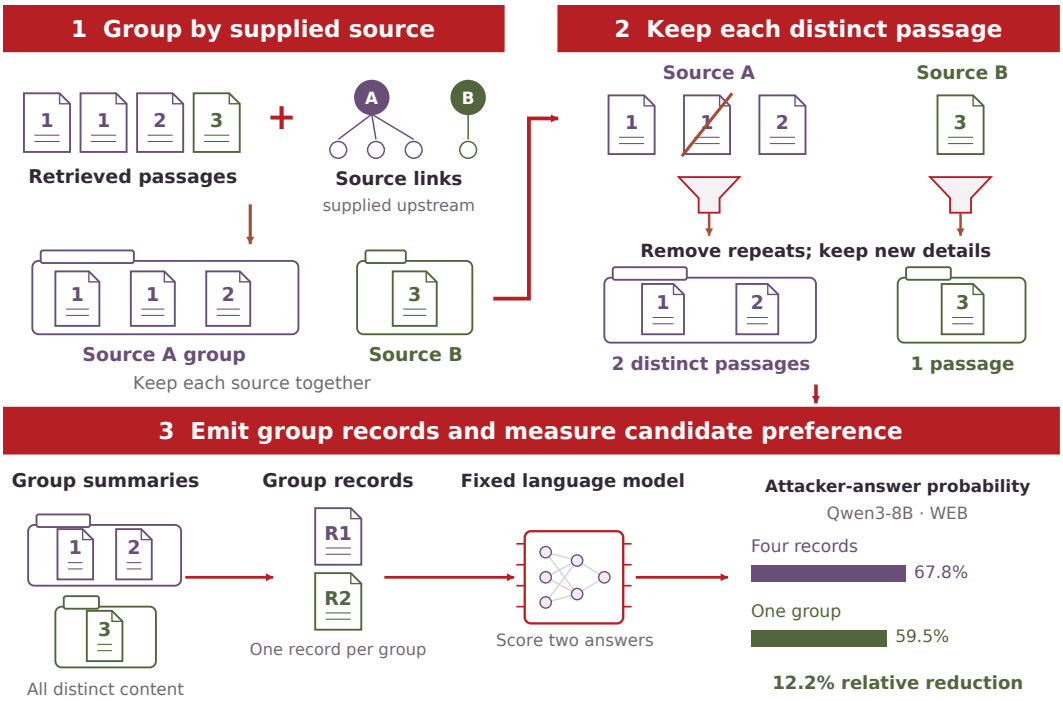


Figure 2: Grouping representation and measured record-layout effect. (1) Supplied links group passages. (2) Canonical content sets remove the repeated passage 1 while retaining complementary passages 2 and 3. (3) The text implementation emits one record per group and scores two candidate answers. Its inset compares six-slot, token-length-matched prompts on Qwen3-8B/WEB: grouping the same 160 words lowers attacker-answer probability from 67.8% to 59.5%, a 12.2% relative reduction. Numerical norm bounds apply separately to the theoretical aggregator.

3.2 Invariant and Faithful Group–Content State

The representation retains the canonical content of every supplied group while removing repeated occurrences within that group.

Figure 2 illustrates how supplied membership determines content retention and emitted records. Its inset shows the six-slot record-layout intervention: on Qwen3-8B/WEB, grouping lowers attacker-answer probability from 67.8% to 59.5%, a 12.2% relative reduction with identical token length.

Let $\mathcal{K}$ be an infinite set of opaque membership labels and let $\mathcal{Z}_{\mathrm{ev}}$ contain canonical evidence elements. A label's spelling carries no evidential meaning; equality and inequality encode group membership. An upstream process supplies the partition, and authenticated provenance, when available, can justify it; the model-facing label remains local bookkeeping. A finite grouped configuration is $x = ((k_i, z_i))_{i=1}^{n} \in \mathsf{Cfg} = (\mathcal{K} \times \mathcal{Z}_{\mathrm{ev}})^*$. Let K_x be its set of occupied labels and, for each $k \in K_x$, let $V_x(k) = \{z_i : k_i = k\}$. The canonical evidence quotient is the finite multiset

$$q_{\mathrm{ev}}(x) = \sum_{k \in K_x} \delta_{V_x(k)} \in \mathsf{Quot} = \mathsf{M}_{\mathrm{fin}}(\mathrm{Fin}^{+}(\mathcal{Z}_{\mathrm{ev}})). \tag{2}$$

The outer object is a multiset, so distinct groups with identical content remain distinct; each inner object is a set, so another canonical copy inside one group disappears. Declare $x \sim y$ when $q_{\mathrm{ev}}(x) = q_{\mathrm{ev}}(y)$. This relation removes record order, bijective renaming of opaque occupied labels, and within-group canonical copies while preserving complementary elements and every nontrivial split or merge as distinct quotient states. Appendix B gives the signature-augmented version that carries authenticated identity or credibility as an evidence-bearing coordinate.

The representation has three nuisance invariances: record permutation, bijective renaming of occupied opaque labels, and within-group insertion or deletion of an exact canonical duplicate while one

occurrence remains. Faithfulness asks that all distinctions between the resulting group–content states remain recoverable.

Theorem 1 (Invariant and faithful evidence representations)**.** *For any set Y and representation $\Psi : \mathsf{Cfg} \to Y$, the following are equivalent: (i) Ψ has the three nuisance invariances above; (ii) $\Psi(x) = \Psi(y)$ whenever $x \sim y$; and (iii) there is a unique $\bar{\Psi} : \mathsf{Quot} \to Y$ such that*

$$\Psi = \bar{\Psi} \circ q_{\mathrm{ev}}. \tag{3}$$

Among these invariant representations, preserving every relation-invariant observable is equivalent to injectivity of $\bar{\Psi}$, and to recoverability of q_{ev} from Ψ. Thus faithful invariant states are precisely injective recodings of q_{ev}.

The three operations generate exactly the fibers of q_{ev}: match groups with equal content sets, rename their labels, delete duplicate occurrences, and permute records. Standard quotient factorization then gives the characterization; choosing q_{ev} itself as the observable gives recoverability. Appendix B supplies the full proofs and the downstream-map formulation.

The group-additive construction is one factorized realization. Let U be an ambient record universe, let $z_{\mathrm{ev}} : U \to \mathcal{Z}_{\mathrm{ev}}$ be one fixed canonicalizer, let $(\mathcal{X}, \|\cdot\|)$ be a real normed vector space, and let a map finite subsets of $\mathcal{Z}_{\mathrm{ev}}$ into $\mathcal{X}$. For a finite $I \subseteq U$ and a partition P of I, define

$$E_I(P) = \sum_{G \in P} a(\{z_{\mathrm{ev}}(i) : i \in G\}). \tag{4}$$

Writing $q_P = \sum_{G \in P} \delta_{\{z_{\mathrm{ev}}(i) : i \in G\}}$ and $\Phi_a(m) = \sum_S m(S)a(S)$ gives $E_I(P) = \Phi_a(q_P)$. Hence E_I always factors through the quotient; quotient faithfulness holds exactly when the aggregator separates the reachable quotient states, as characterized in Appendix B.

Proposition 2 (Equal-count separation)**.** *Let z_1, z_2, z_3, z_4 be distinct canonical elements attached to four fixed records, and set*

$$P_A = \{\{1,2\},\{3,4\}\}, \qquad P_B = \{\{1,3\},\{2,4\}\}.$$

The record content, group count, and block-size multiset agree, but $q_{P_A} \neq q_{P_B}$. For every $B > 0$ there is a fixed scalar aggregator with $|a(S)| \leq B$ for which $E_I(P_A) = 2B$ and $E_I(P_B) = -2B$.

For the witness, assign $+B$ to $\{z_1, z_2\}$ and $\{z_3, z_4\}$, $-B$ to $\{z_1, z_3\}$ and $\{z_2, z_4\}$, and zero elsewhere. Thus the same global content and count can support distinct bounded evidence states. Count records how many units exist; the group–content state also records which facts they combine.

Proposition 3 (Exact-replication invariance)**.** *If a new record $i' \in U \setminus I$ is added to the block containing i and $z_{\mathrm{ev}}(i') = z_{\mathrm{ev}}(i)$, then the evidence state in Equation* (4) *is unchanged.*

The construction uses a supplied grouping key, a fixed canonicalizer, and an extensional set-valued group input; its proof is in Appendix B. A renderer that depends only on this group–content state carries the invariance to the final prompt (Corollary 10). Complementary passages remain separate canonical elements within a group. More generally, let $J \subseteq U \setminus I$ be a finite set of new records that joins an existing block $G \in P$, and set $P' = (P \setminus \{G\}) \cup \{G \cup J\}$. Their entire effect is

$$E_{I \cup J}(P') - E_I(P) = a(A_G \cup A_J) - a(A_G), \tag{5}$$

where $A_G = \{z_{\mathrm{ev}}(i) : i \in G\}$ and $A_J = \{z_{\mathrm{ev}}(j) : j \in J\}$. When $P = P_c^\star$ and every record in J shares the claim-relative root of G, this is same-root proliferation. The number of outer contributions stays fixed; the state changes only through genuinely new canonical content inside the root.

3.3 From Partition Error to Decision Stability

For two partitions P, Q of the same records, define the content-aware discrepancy

$$\Delta_q(P, Q) = \|q_P - q_Q\|_1 = \sum_A |q_P(A) - q_Q(A)|.$$

It cancels matching group–content sets even when their record identities differ. With a fixed canonicalizer and $\|a(A)\| \leq B$, Proposition 12 gives

$$\|E_I(P) - E_I(Q)\| \leq B\Delta_q(P, Q) \leq BH(P, Q), \tag{6}$$

where H counts the blocks in changed overlap components (Appendix B). The first bound is exact in the worst case over bounded scalar aggregators for each fixed pair P, Q. Equal-count reassignment can attain $4B$, as the preceding witness shows. For an L-Lipschitz decision margin m, its sign is stable whenever $|m(E_I(P))| > LB\Delta_q(P, Q)$. Taking $P = P_c^\star$ and $Q = \widehat{P}_c$ connects upstream partition error to representation distortion and then to decision stability.

The numerical guarantee applies to an external aggregator satisfying the cap B and a downstream margin with verified constant L. In the text experiments, the enforced operation is one emitted record per group; the model receives record IDs, stance labels, and text. Stored unit weights are bookkeeping; scoring uses the model's sequence likelihoods. Candidate-probability changes therefore measure the response to record layout. HUMAN keys denote crowd-annotated evidence units; WEB copies share canonical URLs, and its Four-Unit records satisfy frozen domain, URL, and normalized-text separation criteria.

4 Experiments

4.1 Panels, Scoring, and Statistics

The natural-text experiment uses 101 PERSPECTRUM claims in 49 dependence components (Chen et al., 2019) and 138 ConflictingQA questions in 135 components (Wan et al., 2024; Chiang & Lee, 2024) as separate domains. The grouping panel uses 66 HUMAN claims passing a frozen document filter and all 138 WEB questions; generation uses 40 components per domain.

Candidate-choice experiments score two complete, length-matched assistant answers with native termination sequences by total conditional log likelihood. Exact ties score zero in the primary analysis and one half in sensitivity analysis. Controlled generation separately scores the parsed leading `Yes`/`No` over all outputs.

For trial i, let x_i be the rendered prompt, let a_i and b_i be the attack-side and opposite complete candidates, and let $u_{i,c,1:m_{ic}}$ be candidate c's scoring suffix, including the checkpoint's native termination sequence. We compute

$$\begin{aligned} \ell_i(c) &= \sum_{t=1}^{m_{ic}} \log p_\theta(u_{i,c,t} \mid x_i, u_{i,c,<t}), \\ p_{\mathrm{attack},i} &= \frac{\exp \ell_i(a_i)}{\exp \ell_i(a_i) + \exp \ell_i(b_i)} = \frac{1}{1 + \exp(\ell_i(b_i) - \ell_i(a_i))}. \end{aligned} \tag{7}$$

Thus p_{attack} is the likelihood share of the attacker-favored candidate after normalizing the two candidate likelihoods. The attack side identifies the direction targeted by the manipulation; natural-text items include stance questions, so this endpoint measures candidate preference rather than factual accuracy. For example, 71.9% to 45.5% means a 26.4-percentage-point drop and a 36.7% relative reduction. We compute relative reduction as $100(p_{\mathrm{before}} - p_{\mathrm{after}})/p_{\mathrm{before}}$ using model–dataset means. Separately reported strict-selection and generated-answer rates count selection of the targeted answer.

We average mirrored assignments and orders within item before each contrast. The relative summaries reuse these paired predictions and divide aggregate means; they are computed after the original analysis. Their intervals resample both arms together, so uncertainty in the denominator is retained. Fixed-seed 10,000-replicate bootstraps resample items or natural-text dependence components; estimates remain checkpoint- and domain-specific. Appendix C.4–C.5 gives full specifications.

4.2 Central Grouping Intervention and Breadth Studies

For the one-group partial-copy panel, four non-overlapping 40-word windows come from one frozen document. For the four-group panel, four windows come from four operationally distinct supplied units. We hold the supplied keys fixed, inject false splits or false merges downstream, aggregate content within resulting groups, and emit one evidence record per estimated group. The main comparisons hold the ordered 160 attack words fixed. Across Qwen3-8B and Qwen3-4B (Yang et al., 2025), Phi-4-mini (Microsoft, 2025), and Mistral-7B (Jiang et al., 2023), this experiment contributes 39,168 trials. A further 13,056-trial control uses six JSON slots in both arms, retains R1–R6 and

Table 1: Attacker-answer probability before and after correct same-root grouping. The same ordered 160 evidence words are presented as four falsely split records (before) or one full-content group (after). Relative reduction is $100(p_{\text{before}} - p_{\text{after}})/p_{\text{before}}$. Intervals are pointwise 95% paired dependence-component bootstrap intervals for the relative reduction; HUMAN has 66 claims in 32 components, and WEB has 138 claims in 135 components.

Model	Data	Before (four records)	After (one group)	Relative reduction (95% interval)
Qwen3-8B	HUMAN	71.9%	45.5%	36.7% [29.2, 44.4]
Qwen3-8B	WEB	68.6%	43.3%	36.8% [32.0, 41.5]
Qwen3-4B	HUMAN	65.8%	45.1%	31.3% [23.8, 38.7]
Qwen3-4B	WEB	63.4%	39.0%	38.5% [33.7, 43.6]
Phi-4-mini	HUMAN	63.2%	48.7%	22.9% [16.7, 29.3]
Phi-4-mini	WEB	56.4%	46.1%	18.2% [13.9, 22.8]
Mistral-7B	HUMAN	80.2%	47.6%	40.7% [36.1, 46.6]
Mistral-7B	WEB	74.2%	48.9%	34.0% [30.1, 37.9]

every header and separator, matches final character, byte, and tokenizer lengths, and varies whether the fixed word stream occupies one or four nonempty content-bearing slots.

Controlled campaign and independent-root panel. We construct 48 fictional product pairs. For each attack side, four 40-word cross-genre records deterministically restate one frozen three-fact campaign brief and share one oracle root; a second family contains four 40-word records tied to independently specified laboratory, panel, endurance, and service-data roots. Both families use the matched six-slot renderer with identical ordered words and token length; Appendix C.5 gives the atomic-fact and serialization details. Two attack sides and two order mirrors yield 3,072 choice trials over the same four checkpoints.

Additional experiments cover authority cues, explicit relation/count stages, repetition dose, natural-text copies, added Qwen3-14B-AWQ and Phi-4 checkpoints (Yang et al., 2025; Abdin et al., 2024), quantization, and controlled generation. Together with the central grouping, matched-serialization, and provenance panels, these controlled studies contain 104,402 trials across 6 unique checkpoints. Appendix E lists inference settings and evaluation counts.

Real-question evaluation with estimated document groups. On 120 document-disjoint HotpotQA questions (Yang et al., 2018), Mistral-7B and Phi-4 answer contexts with four added exact copies or overlapping excerpts from a distractor document. Raw, exact-hash, MiniLM, and visible-title grouping preserve all unique sentences. Official answer EM/F1 and paired question resampling measure correctness (Appendix F).

5 Results

5.1 Reduction in Attacker-Answer Probability on Natural-Text Evidence

Table 1 shows the candidate-preference effect: with the same 160 evidence words, one full-content group reduces attacker-answer probability by 18.2–40.7% relative to four falsely split records. Qwen3-8B falls from 71.9% to 45.5% on HUMAN and from 68.6% to 43.3% on WEB, relative reductions of 36.7% and 36.8%. Every model–dataset pair has a positive reduction with a pointwise 95% interval above zero. The three-contrast decomposition in Appendix Table 6 separates this effect from content retention and false merging.

Complementary passages still contribute inside one group. The full-content arm exact-deduplicates and concatenates four non-overlapping windows into one 160-word group record. For Qwen3-8B, retaining all four windows instead of one raises attacker-answer probability from 31.5% to 45.5% on HUMAN (44.3% relative) and from 29.0% to 43.3% on WEB (49.2% relative). Across all eight pairs, the additional complementary content contributes 2.91–29.82 percentage

Table 2: Controlled campaign and independent-root interventions. Values are percentage-point changes in p_{attack} with pointwise 95% dependence-component bootstrap intervals. Same-root split is four content-bearing records minus one grouped record; independent-root merge loss is four records minus one falsely merged record. Positive values are the preregistered directions; checkpoints are reported without pooling.

Model	Same-root split	Independent-root merge loss
Qwen3-8B	$+18.46[+15.55, +21.41]$	$-4.25[-6.86, -1.87]$
Qwen3-4B	$+9.66[+6.59, +12.54]$	$+5.39[+3.12, +7.94]$
Phi-4-mini	$+27.44[+24.64, +30.25]$	$+15.73[+12.87, +18.70]$
Mistral-7B	$-4.44[-7.35, -1.64]$	$+3.02[+1.63, +4.61]$

points, with all pointwise intervals above zero. This is the influence of retained content; the same-root grouping comparison above then removes the extra influence of presenting those words as four records.

Keeping independent roots separate preserves their influence. For Qwen3-8B/HUMAN, four independent supplied units receive 70.6% attacker-answer probability, compared with 45.4% after a false merge. Restoring their separate groups recovers 55.4% relative to the falsely merged baseline; the corresponding WEB increase is 56.6%. Across all model–dataset pairs, false merging removes 9.13–31.79 percentage points, while false splitting one supplied unit adds 10.27–32.66 points. Every pointwise interval is positive. Both comparisons keep the evidence words and their order identical. In this panel, correct grouping suppresses same-root multiplicity while retaining complementary content and distinct-root contributions.

The effect persists when record fields and token length are matched. In the stricter six-slot experiment, same-root grouping reduces attacker-answer probability by 1.2–15.6% across eight model–dataset pairs; six of their eight pointwise 95% intervals lie above zero. For Qwen3-8B/WEB, the probability falls from 67.8% to 59.5% (12.2% relative; Figure 2). Including the eight independent-root merge contrasts gives 16/16 positive estimates, ranging from $+0.63$ to $+13.33$ points, with 12 of 16 intervals above zero. This smaller effect measures content placement under equal object, header, and token counts. The two rendering designs estimate distinct interventions. Appendix C.5 gives all estimates.

Post-hoc Bonferroni familywise 95% bootstrap intervals retain positive lower bounds for all 16 content-fixed split/merge effects, seven of eight content-retention effects, and 11 of 16 matched six-slot contrasts (Appendix D.5). The remaining content-retention interval, for Phi-4-mini/WEB, crosses zero by 0.02 points.

5.2 Controlled GEO Renderings Expose Checkpoint-Dependent Partition Effects

Correctly grouping the four same-root campaign records lowers attacker-answer probability from 77.8% to 59.3% for Qwen3-8B (23.7% relative), from 89.9% to 80.3% for Qwen3-4B (10.7%), and from 77.4% to 50.0% for Phi-4-mini (35.4%). Mistral-7B reverses direction: grouping raises the probability from 53.1% to 57.6% (8.4%). Table 2 retains the original percentage-point contrasts and their intervals. Coordinated proliferation can therefore gain influence without exact copying, and its behavioral direction depends on the model.

The independent-root merge loss is positive for Qwen3-4B, Phi-4-mini, and Mistral-7B but reverses for Qwen3-8B. The balanced average is a boundary-sensitivity stress test: mirror decomposition shows large order interactions (for Qwen3-8B, the independent-root contrast is $+41.43$ points in original order and -49.92 in reverse order). Record-boundary placement is therefore behaviorally active, with its sign determined jointly by checkpoint, content family, and presentation.

Appendix Table 7 decomposes these effects by order.

5.3 Record Copies Reweight Candidate and Generated Decisions

Four copies can change roughly 9–52 additional choices per 100 decisions toward the attacker-favored answer. Across six checkpoints, the measured shifts are 9.24–51.98 percentage points; every model–domain interval excludes replication invariance in the observed direction (Figure 5 and Appendix Tables 13 and 8).

The original HUMAN and WEB ranges are 14.11–49.01 and 9.24–37.14 points. Added 14B checkpoints reach 30.62–51.98 points. In controlled answer-plus-one-sentence generation, copying produces roughly 16–21 additional attacker-favored leading answers per 100 generations (15.62–20.62 percentage points across three checkpoints), with positive lower bounds and 2,880/2,880 valid outputs. Appendix C.5 reports prompt-rule effects and threshold sensitivity.

Grouping improves generated answers under overlapping excerpts. Fixed MiniLM grouping raises answer EM from 20.00% to 25.42% for Mistral-7B and from 28.75% to 34.58% for Phi-4: paired gains of 5.42 and 5.83 points, with pointwise 95% intervals $[2.08, 9.17]$ and $[1.67, 10.00]$. Exact-copy gains are 4.17 points for Mistral ($[0.83, 7.50]$) and 0.83 for Phi-4 ($[-2.92, 5.00]$). Tables 17–18 report both conditions and all grouping comparators. This retrospective reporting subset tests controlled distractor proliferation; document identity defines the reference groups, and natural upstream ancestry remains unknown.

6 Discussion

Partition is an evidence-accounting decision. Content-fixed comparisons show that grouping changes measured influence while retaining complementary content. The stricter six-slot estimates are smaller, with six of eight same-root intervals above zero; campaign effects reverse across checkpoints and orders (Table 2). Record partition is therefore a behavioral variable whose effect depends on the model and presentation. Evidence roots describe claim-relevant provenance; surface diversity and URL count describe observable records.

Dependence-aware weighting belongs before generation. A source-aware system separates partition estimation, content retention, and numerical aggregation. On the fictional panel, exact hash and MinHash miss cross-genre same-root groups, while fixed-threshold embeddings (Reimers & Gurevych, 2019) recover both structures (Appendix Table 4). Those predicted partitions coincide with executed arms, so their downstream comparison reuses existing endpoints. The HotpotQA extension executes estimated partitions on real contexts; its reference relation is document identity, with natural information-source ancestry unknown.

What information must survive grouping. Theorem 1 characterizes faithful retention of an invariant group–content state by injective recoding. Proposition 2 shows what scalar group counts discard, even with identical records and equal block sizes. The discrepancy Δ_q carries this distinction into error propagation by charging unmatched group–content mass.

7 Conclusion

A supplied partition specifies which content contributes together, preserving distinctions beyond record and group counts. With all 160 evidence words retained, grouping reduces attacker-answer probability by 18.2–40.7%; the stricter six-slot control yields 1.2–15.6% reductions. Campaign effects reverse across checkpoints and orders. The theory bounds partition-error distortion for a fixed, bounded aggregator. The text implementation enforces deduplication and one-record emission. On real questions with overlapping distractor excerpts, estimated grouping improves generated-answer accuracy for Mistral-7B and Phi-4.

AI use statement

Generative AI tools were used to aid and polish the writing of this paper, including revisions to wording, clarity, and organization. The authors take responsibility for the final content of the paper.

Ethics statement

The study analyzes public, previously released datasets, including prior crowd annotations in the HUMAN domain, and introduces zero new human-subject interactions. All analyzed data are public. The main misuse risk is content manipulation informed by the controlled repetition protocol. The presentation centers defensive evaluation and separates oracle source metadata from editable text labels.

Reproducibility statement

Supplementary materials include prompts, predictions, grouping assignments, and analysis scripts for the controlled panels and real-question extension. The scripts recompute paired contrasts and official answer scores. The reported extension contains 3,840 aligned conditions supported by 2,644 distinct model inputs; each question remains one statistical unit.

References

Marah Abdin et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. URL https://arxiv.org/abs/2412.08905.

Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, and Ameet Deshpande. GEO: Generative engine optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5–16, 2024. doi: 10.1145/3637528.3671900. URL https://doi.org/10.1145/3637528.3671900.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. Seeing things from a different angle: Discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 542–557, 2019. doi: 10.18653/v1/N19-1053. URL https://aclanthology.org/N19-1053/.

Yimeng Chen, Zhe Ren, Firas Laakom, Yu Li, Dandan Guo, and Jürgen Schmidhuber. How much can we trust LLM search agents? measuring endorsement vulnerability to web content manipulation. *arXiv preprint arXiv:2606.16821*, 2026. doi: 10.48550/arXiv.2606.16821. URL https://arxiv.org/abs/2606.16821.

Cheng-Han Chiang and Hung-yi Lee. Do metadata and appearance of the retrieved webpages affect LLM's reasoning in retrieval-augmented generation? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 389–406, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.24. URL https://aclanthology.org/2024.blackboxnlp-1.24/.

Ziyu Ge, Yuhao Wu, Daniel Wai Kit Chin, Roy Ka-Wei Lee, and Rui Cao. Resolving conflicting evidence in automated fact-checking: A study on retrieval-augmented LLMs. *arXiv preprint arXiv:2505.17762*, 2025. URL https://arxiv.org/abs/2505.17762.

Albert Q. Jiang et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16867–16878, 2024. URL https://aclanthology.org/2024.lrec-main.1466/.

Yuxuan Li, Xinwei Guo, Jiashi Gao, Guanhua Chen, Xiangyu Zhao, Jiaxin Zhang, Quanying Liu, Haiyan Wu, Xin Yao, and Xuetao Wei. LLMs trust humans more, that's a problem! unveiling and mitigating the authority bias in retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28844–28858, 2025. doi: 10.18653/v1/2025.acl-long.1400. URL https://aclanthology.org/2025.acl-long.1400/.

Junchi Liao. Auditing provenance sensitivity in LLM agent action selection. *arXiv preprint arXiv:2607.20827*, 2026. URL https://arxiv.org/abs/2607.20827.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7052–7063, 2021. doi: 10.18653/v1/2021.emnlp-main.565. URL https://aclanthology.org/2021.emnlp-main.565/.

Microsoft. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025. URL https://arxiv.org/abs/2503.01743.

Atharv Naphade. Rational synthesizers or heuristic followers? analyzing LLMs in RAG-based question-answering. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 40293–40311, 2026. doi: 10.18653/v1/2026.findings-acl.2003. URL https://aclanthology.org/2026.findings-acl.2003/.

Irwan Rahadi. Counting copies as evidence: Confidence inflation from dependent evidence in retrieval-augmented generation (RAG), August 2026. URL https://doi.org/10.5281/zenodo.21923648. Position paper and preprint.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Jonathan J. Ross, Bevan Koopman, Anton van der Vegt, and Guido Zuccon. How retriever redundancy and diversity impact RAG effectiveness. *arXiv preprint arXiv:2608.13956*, 2026. URL https://arxiv.org/abs/2608.13956.

Jakob Schuster, Vagrant Gautam, and Katja Markert. Whose facts win? LLM source preferences under knowledge conflicts. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29430–29459, 2026. doi: 10.18653/v1/2026.acl-long.1357. URL https://aclanthology.org/2026.acl-long.1357/.

Dachuan Song, Junyu Yin, Zechen Hu, and Xuan Wang. Mergeable model-side aggregation states for long-context language models. *arXiv preprint arXiv:2607.26448*, 2026. URL https://arxiv.org/abs/2607.26448.

Kaiser Sun, Fan Bai, and Mark Dredze. Task matters: Knowledge requirements shape LLM responses to context–memory conflict. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4154–4176, 2026. doi: 10.18653/v1/2026.findings-acl.202. URL https://aclanthology.org/2026.findings-acl.202/.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024. URL https://arxiv.org/abs/2404.13208.

Alexander Wan, Eric Wallace, and Dan Klein. What evidence do language models find convincing? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7468–7484, 2024. doi: 10.18653/v1/2024.acl-long.403. URL https://aclanthology.org/2024.acl-long.403/.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Zhonghao Zhan, Huichi Zhou, Zhenhao Li, Peiyuan Jing, Krinos Li, and Hamed Haddadi. How adversarial environments mislead agentic AI? In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10264–10280, 2026. doi: 10.18653/v1/2026.findings-acl.499. URL https://aclanthology.org/2026.findings-acl.499/.

Sizhe Zhou, Sha Li, Yu Meng, Yizhu Jiao, Heng Ji, and Jiawei Han. Establishing knowledge preference in language models. *arXiv preprint arXiv:2407.13048*, 2024. URL https://arxiv.org/abs/2407.13048.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14544–14556, 2023. doi: 10.18653/v1/2023.findings-emnlp.968. URL https://aclanthology.org/2023.findings-emnlp.968/.

A TWO-STAGE RESPONSE THEORY

A.1 FINITE-STATE TWO-STAGE IDENTITY

For item i, let balanced direction $S_i \in \{-1, +1\}$ determine which abstract value wins the supplied-group count. Let $Z_i^{\text{parse}} \in \mathcal{Z}$ be an explicit parse-stage output and $\mu_i(z)$ the mean response of a separate use-stage call when a program supplies state z. Define

$$\pi_i(z) = \frac{1}{2}\left[\Pr(Z_i^{\text{parse}} = z \mid S_i = +1, i) - \Pr(Z_i^{\text{parse}} = z \mid S_i = -1, i)\right].$$

Theorem 4 (Finite-state two-stage identity)**.** *If the executed two-stage program satisfies* $\mathbb{E}[Y_i^{\text{cas}} \mid Z_i^{\text{parse}} = z, S_i = s, i] = \mu_i(z)$*, then its half contrast is*

$$\begin{aligned} r_i^{\text{cas}} &= \frac{\mathbb{E}[Y_i^{\text{cas}} \mid S_i = +1, i] - \mathbb{E}[Y_i^{\text{cas}} \mid S_i = -1, i]}{2} \\ &= \sum_{z \in \mathcal{Z}} \mu_i(z)\pi_i(z). \end{aligned} \tag{8}$$

Proof of Theorem 4. For each $s \in \{-1, +1\}$, the law of total expectation and the explicit two-stage condition give

$$\mathbb{E}[Y_i^{\text{cas}} \mid S_i = s, i] = \sum_{z \in \mathcal{Z}} \mu_i(z) \Pr(Z_i^{\text{parse}} = z \mid S_i = s, i).$$

Subtracting the two directions and dividing by two gives Equation (8). The derivation accommodates dependent stages and direction-asymmetric errors. □

For a binary parse state, any use response has the form $\mu_i(z) = a_i + u_i z$. If

$$p_i = \frac{\mathbb{E}[Z_i^{\text{parse}} \mid S_i = +1, i] - \mathbb{E}[Z_i^{\text{parse}} \mid S_i = -1, i]}{2},$$

then $r_i^{\text{cas}} = p_i u_i$. Across items, $\mathbb{E}[p_i u_i] = \mathbb{E}[p_i]\mathbb{E}[u_i] + \text{Cov}(p_i, u_i)$. Exact reconstruction from pooled parse and use means therefore requires control of the item-level covariance.

A.2 TWO-WORLD TRANSPORT LOWER BOUND

Proposition 5 (Prompt-transport lower bound)**.** *For responses bounded in* $[-L, L]$ *under arbitrary relationships between auxiliary and end-to-end prompts, two end-to-end worlds can have identical auxiliary distributions while their end-to-end effects are opposite. Every predictor determined by the auxiliary distribution then has worst-case absolute error at least* L*.*

Proof of Proposition 5. Fix any auxiliary-task distribution and any predictor $\widehat{R}$ measurable with respect to it. Construct two compatible end-to-end worlds with the same auxiliary observations:

$$\mathcal{W}_+ : Y^{\text{e2e}} = LS, \qquad \mathcal{W}_- : Y^{\text{e2e}} = -LS.$$

Their end-to-end effects are $+L$ and $-L$. The predictor has the same value in both worlds, while

$$2L \le |\widehat{R} - L| + |\widehat{R} + L|.$$

At least one error is at least L. □

A.3 COMMON-STATE THREE-DEFECT BOUND

Let Z^{e2e} be a reference end-to-end state on the same finite state space, let $h_i(z)$ be its reference downstream response, and define

$$R_{\text{e2e}} = \mathbb{E}[S_i Y_i^{\text{e2e}}], \qquad R_{\text{cas}} = \mathbb{E}[S_i \mu_i(Z_i^{\text{parse}})].$$

Introduce

$$\eta_{\mathrm{dir}} = \left|\mathbb{E}[S_i\{Y_i^{\mathrm{e2e}} - h_i(Z_i^{\mathrm{e2e}})\}]\right|,$$
$$\eta_{\mathrm{use}} = \mathbb{E}\left[|h_i(Z_i^{\mathrm{e2e}}) - \mu_i(Z_i^{\mathrm{e2e}})|\right],$$
$$\eta_{\mathrm{parse}} = \mathbb{E}\left[|\mu_i(Z_i^{\mathrm{e2e}}) - \mu_i(Z_i^{\mathrm{parse}})|\right].$$

Adding and subtracting the two intermediate responses and applying the triangle inequality gives

$$|R_{\mathrm{e2e}} - R_{\mathrm{cas}}| \le \eta_{\mathrm{dir}} + \eta_{\mathrm{use}} + \eta_{\mathrm{parse}}.$$

On a common reference state space, the transport gap decomposes into the three displayed defects. If direct mismatch is zero, use mismatch is uniformly at most ϵ_u, and the parse states disagree with probability at most ϵ_p, with responses in $[-L, L]$, then

$$|R_{\mathrm{e2e}} - R_{\mathrm{cas}}| \le \epsilon_u + 2L\epsilon_p.$$

B Canonical Evidence Quotient and External Grouping Guarantees

B.1 Text-Only Provenance Recovery Boundary

Proposition 6 (Exact boundary for deterministic text-only provenance recovery)**.** *Fix a claim c and an integer $n \ge 2$. Let Π_n be the set of partitions of $[n]$ and let $\mathcal{W}_{c,n}$ be the class of admissible worlds with n observed records. Let $Z_c : \mathcal{W}_{c,n} \to \mathcal{Z}^n$ map each world to its complete model-visible text observation, and let $P_c^\star : \mathcal{W}_{c,n} \to \Pi_n$ map each world to its oracle claim-relative provenance partition. There exists a deterministic text-only rule $\Gamma_c : \mathcal{Z}^n \to \Pi_n$ satisfying $\Gamma_c(Z_c(w)) = P_c^\star(w)$ for every $w \in \mathcal{W}_{c,n}$ if and only if, for all $w_0, w_1 \in \mathcal{W}_{c,n}$,*

$$Z_c(w_0) = Z_c(w_1) \quad \Longrightarrow \quad P_c^\star(w_0) = P_c^\star(w_1).$$

Proof of Proposition 6. For necessity, suppose a universally exact rule Γ_c exists. If $Z_c(w_0) = Z_c(w_1)$, then functionality gives $\Gamma_c(Z_c(w_0)) = \Gamma_c(Z_c(w_1))$, while exactness identifies the two sides with $P_c^\star(w_0)$ and $P_c^\star(w_1)$. Hence $P_c^\star(w_0) = P_c^\star(w_1)$.

For sufficiency, suppose $P_c^\star$ is constant on each observational fiber of Z_c. For every $z \in Z_c(\mathcal{W}_{c,n})$, define $\Gamma_c(z)$ as the unique oracle-partition value attained on the fiber $Z_c^{-1}(\{z\})$. Fiber constancy makes this definition well-defined. Since Π_n is nonempty, extend Γ_c arbitrarily outside $Z_c(\mathcal{W}_{c,n})$. Then $\Gamma_c(Z_c(w)) = P_c^\star(w)$ for every $w \in \mathcal{W}_{c,n}$. □

B.2 Operational Quotient and Factorization

Let $\mathsf{Cont} = \mathrm{Fin}^+(\mathcal{Z}_{\mathrm{ev}})$ be the nonempty finite canonical-content sets and let

$$\mathsf{Quot} = \mathsf{M}_{\mathrm{fin}}(\mathsf{Cont}) = \{m : \mathsf{Cont} \to \mathbb{N}_0 : \mathrm{supp}(m) \text{ is finite}\}.$$

For $x = ((k_i, z_i))_{i=1}^n \in \mathsf{Cfg}$, write K_x for its occupied labels and $V_x(k) = \{z_i : k_i = k\}$. Equation (2) maps x to the outer multiset of these inner sets. Label equality encodes membership; label spelling is removed by the quotient.

Identity-sensitive representations use an explicit authenticated signature coordinate. Let Σ be a nonempty space of evidence-bearing group signatures and define

$$\mathsf{Cfg}_\Sigma = \{(x, s_x) : x \in \mathsf{Cfg},\ s_x : K_x \to \Sigma\}, \qquad \mathsf{Quot}_\Sigma = \mathsf{M}_{\mathrm{fin}}(\Sigma \times \mathsf{Cont}),$$

$$q_{\mathrm{ev}}^\Sigma(x, s_x) = \sum_{k \in K_x} \delta_{(s_x(k), V_x(k))}. \tag{9}$$

Declare $(x, s_x) \sim_\Sigma (y, s_y)$ exactly when $q_{\mathrm{ev}}^\Sigma(x, s_x) = q_{\mathrm{ev}}^\Sigma(y, s_y)$. A bijective label renaming $\beta : K_x \to K'$ sends $((k_i, z_i)_i, s_x)$ to $((\beta(k_i), z_i)_i, s')$, where $s'(\beta(k)) = s_x(k)$. Thus label spelling is removed while each signature–content pair is retained. Because $\mathcal{K}$ is infinite, q_{ev}^Σ is surjective. Consequently, for every $\Psi_\Sigma : \mathsf{Cfg}_\Sigma \to Y$, invariance on $\sim_\Sigma$ classes is equivalent to a unique

factorization $\Psi_\Sigma = \bar{\Psi}_\Sigma \circ q_{\mathrm{ev}}^{\Sigma}$. The sufficiency, faithfulness, and prompt-visible statements below have the same typed analogues. For $a : \Sigma \times \mathsf{Cont} \to \mathcal{X}$, the additive analogue is

$$E_a^{\Sigma}(x, s_x) = \sum_{k \in K_x} a(s_x(k), V_x(k)), \qquad \Phi_a^{\Sigma}(m) = \sum_{(\sigma, A)} m(\sigma, A) a(\sigma, A),$$

and its faithfulness criterion is finite integer-linear independence over $(\sigma, A) \in \Sigma \times \mathsf{Cont}$.

Proposition 7 (Operational quotient)**.** *The map $q_{\mathrm{ev}} : \mathsf{Cfg} \to \mathsf{Quot}$ is surjective. Equality of its outputs is exactly the equivalence relation generated by record permutation, bijective renaming of occupied opaque labels, and insertion or deletion of a within-group canonical duplicate while one occurrence remains. A nontrivial split of one occupied group into $r \geq 2$ nonempty groups, or a nontrivial merge of $r \geq 2$ occupied groups, changes the quotient state even when the union of visible canonical content is unchanged.*

Proof. Each declared nuisance transformation leaves the outer multiset of inner content sets unchanged. Conversely, if $q_{\mathrm{ev}}(x) = q_{\mathrm{ev}}(y)$, equality of finite multisets gives a bijection between occupied labels whose matched groups have identical content sets. Rename labels by that bijection, delete duplicate occurrences inside each matched group, and permute the remaining records; the reduced configurations coincide. Reversing duplicate deletions recovers the originals, so the stated moves generate the whole equivalence relation.

For surjectivity, write any $m \in \mathsf{Quot}$ as $m = \sum_{j=1}^{r} \delta_{A_j}$. Choose r distinct labels and list one pair (k_j, z) for each $z \in A_j$; the resulting configuration maps to m. The empty configuration maps to the empty multiset. Finally define $N(m) = \sum_A m(A)$. A split changes N by $r - 1$ and a merge by $1 - r$; both operations therefore change the quotient state. □

Proof of Theorem 1. Proposition 7 identifies the three-operation invariance with constancy on quotient fibers. If $\Psi = \bar{\Psi} \circ q_{\mathrm{ev}}$ and $x \sim y$, equality of the quotient states gives $\Psi(x) = \Psi(y)$. Conversely, suppose Ψ is invariant. For $m \in \mathsf{Quot}$ choose any x_m with $q_{\mathrm{ev}}(x_m) = m$ and set $\bar{\Psi}(m) = \Psi(x_m)$. Invariance makes this definition independent of the representative, and $\Psi = \bar{\Psi} \circ q_{\mathrm{ev}}$. Surjectivity of q_{ev} makes $\bar{\Psi}$ unique.

If Ψ is invariant, every composition $h \circ \Psi$ is invariant. Conversely, quantifying over all downstream maps includes the identity map on Y; equivalently, binary maps can separate any two unequal points of Y. For a specified downstream family, point separation on $\Psi(\mathsf{Cfg})$ supplies the converse. The final faithfulness and recoverability equivalences follow from Proposition 8 below, whose proof uses only the factorization just established. □

B.3 Faithfulness, Minimal Sufficiency, and Additive Realizations

An enforcing representation $\Theta : \mathsf{Cfg} \to W$ is *quotient-faithful* when $\Theta(x) = \Theta(y)$ if and only if $x \sim y$. It is *universally sufficient* when, for every set D, every relation-invariant observable $f : \mathsf{Cfg} \to D$ factors through Θ. Together, quotient faithfulness and universal sufficiency characterize exact preservation of relation-invariant information.

Proposition 8 (Minimal sufficiency and faithful recoding)**.** *The quotient q_{ev} is universally sufficient. If Θ is universally sufficient, then $q_{\mathrm{ev}} = \rho \circ \Theta$ for some $\rho : W \to \mathsf{Quot}$. If $\Theta = \bar{\Theta} \circ q_{\mathrm{ev}}$ is enforcing, then the following are equivalent: Θ is universally sufficient; Θ is quotient-faithful; and $\bar{\Theta}$ is injective on* Quot.

Proof. Theorem 1 factors every invariant f through q_{ev}, proving its universal sufficiency. If Θ is universally sufficient, apply its definition to the invariant observable $f = q_{\mathrm{ev}}$ to obtain $q_{\mathrm{ev}} = \rho \circ \Theta$. Hence $\Theta(x) = \Theta(y)$ implies $q_{\mathrm{ev}}(x) = q_{\mathrm{ev}}(y)$. Enforcement supplies the reverse implication, proving faithfulness.

For an enforcing $\Theta = \bar{\Theta} \circ q_{\mathrm{ev}}$, faithfulness implies injectivity of $\bar{\Theta}$ by choosing representatives of any two quotient states. Conversely, injectivity of $\bar{\Theta}$ makes equality of Θ equivalent to equality of q_{ev}. The inverse of $\bar{\Theta}$ on the reachable image recovers q_{ev}, so every invariant observable factors through Θ; extend that factor arbitrarily outside the reachable image. □

Let $a : \mathsf{Cont} \to \mathcal{X}$ be a deterministic group aggregator and define

$$E_a(x) = \sum_{k \in K_x} a(V_x(k)), \qquad \Phi_a(m) = \sum_{A \in \mathsf{Cont}} m(A)a(A). \tag{10}$$

Proposition 9 (Additive faithfulness criterion). *Every additive realization satisfies $E_a = \Phi_a \circ q_{\mathrm{ev}}$ and therefore enforces the declared nuisance invariance. It is quotient-faithful if and only if the family $\{a(A) : A \in \mathsf{Cont}\}$ is finitely integer-linearly independent:*

$$\sum_{A \in \mathsf{Cont}} c(A)a(A) = 0, \qquad c \in \mathbb{Z}^{(\mathsf{Cont})} \quad \Longrightarrow \quad c = 0,$$

where the parenthesized exponent denotes finite support. A faithful realization always exists by taking $\mathcal{X} = \mathbb{R}^{(\mathsf{Cont})}$ and $a(A) = e_A$, the canonical coordinate vector.

Proof. Grouping equal outer multiset atoms in Equation (10) gives $E_a(x) = \Phi_a(q_{\mathrm{ev}}(x))$. If a nonzero integer relation exists, decompose $c = c^+ - c^-$ into distinct nonnegative finite multisets. Then $\Phi_a(c^+) = \Phi_a(c^-)$, and surjectivity supplies two distinct quotient states that collide under E_a. Conversely, any collision $\Phi_a(m) = \Phi_a(m')$ for $m \neq m'$ yields the nonzero integer relation $c = m - m'$. The canonical coordinate vectors are integer-linearly independent and expose every multiset coefficient, so they give a faithful realization. □

Corollary 10 (Prompt-visible enforcement criterion). *Let $R : \mathsf{Cfg} \to \mathcal{P}$ be the actual model-visible renderer. Uniform invariance of $h \circ R$ for every binary downstream map $h : \mathcal{P} \to \{0, 1\}$ holds if and only if $R = \bar{R} \circ q_{\mathrm{ev}}$. Therefore, if $x \sim y$ but $R(x) \neq R(y)$, a deterministic binary downstream rule exists that distinguishes the two rendered prompts.*

Proof. Apply Theorem 1 to R. Any pair of distinct rendered prompts is separated by a binary map, establishing necessity of the factorization criterion. □

The corollary characterizes uniform invariance over all binary downstream maps; fixed-model response is measured empirically. The default quotient carries content and group multiplicity, while Equation (9) additionally preserves authenticated identity, credibility, or reliability metadata. Opaque membership labels encode membership in both versions.

B.4 Concrete Additive Construction and Grouping-Error Bound

Let record r support target T or competitor C, and let $g(r)$ denote its supplied group. With group sets G_T and G_C, a unit-weight reference margin is

$$D = b + |G_T| - |G_C|, \tag{11}$$

where b is the pre-context preference for T. Copying a record while preserving the group set leaves this reference margin unchanged. In Equation (4), each group's canonical content set supplies one summand; imposing $\|a(S)\| \leq B$ bounds that summand. The text experiments implement a distinct, structural operation: stable representative selection or exact deduplication followed by within-group concatenation, with one emitted record per group. The stored `unit_weight` value is bookkeeping; the model sees record IDs, stance labels, and text, and its candidate scores come from sequence likelihoods. Our mirrored topology gives the favored side two groups and the other side one group, then reverses that assignment while holding the records fixed.

We now give the complete guarantees for the external representation in Equation (4). The ambient record universe U, canonicalizer $z_{\mathrm{ev}} : U \to \mathcal{Z}_{\mathrm{ev}}$, and deterministic aggregator a are fixed across all compared partitions, and a receives the extensional set of canonical group elements.

Proof of Proposition 3. Let $G \in P$ be the block containing i, set $I' = I \cup \{i'\}$, and define

$$P' = (P \setminus \{G\}) \cup \{G \cup \{i'\}\}.$$

Because $z_{\mathrm{ev}}(i') = z_{\mathrm{ev}}(i)$,

$$\{z_{\mathrm{ev}}(j) : j \in G \cup \{i'\}\} = \{z_{\mathrm{ev}}(j) : j \in G\}.$$

The changed block therefore supplies exactly the same set to the fixed aggregator a, while every other block is unchanged. The two finite sums agree term by term, so $E_{I'}(P') = E_I(P)$. □

For two partitions P and Q of the same finite record set I, build a bipartite overlap graph whose vertices are their blocks and whose edges join blocks with nonempty record intersection. Exclude exactly those connected components containing one P-block and one Q-block with identical record sets. Let $\mathcal{C}_{\mathrm{err}}$ denote the remaining components; for $c \in \mathcal{C}_{\mathrm{err}}$, let p_c and q_c be its numbers of P- and Q-blocks, and define

$$H(P, Q) = \sum_{c \in \mathcal{C}_{\mathrm{err}}} (p_c + q_c).$$

Proposition 11 (Partition-error bound)**.** *Suppose $B \geq 0$ and $\|a(S)\| \leq B$ for every semantic set that occurs under P or Q. Then*

$$\|E_I(P) - E_I(Q)\| \leq BH(P, Q).$$

Proof. Every excluded component contributes the same record block under both partitions. The same z_{ev} and a therefore produce identical vectors, which cancel. In an erroneous component c, write P_c and Q_c for its blocks. The triangle inequality and the assumed bound give

$$\begin{aligned}\left\| \sum_{G \in P_c} a(\{z_{\mathrm{ev}}(i) : i \in G\}) - \sum_{G \in Q_c} a(\{z_{\mathrm{ev}}(i) : i \in G\}) \right\| &\leq \sum_{G \in P_c} \|a(\{z_{\mathrm{ev}}(i) : i \in G\})\| \\ &\quad + \sum_{G \in Q_c} \|a(\{z_{\mathrm{ev}}(i) : i \in G\})\| \\ &\leq B(p_c + q_c).\end{aligned}$$

The erroneous components are disjoint. Summing their differences and applying the triangle inequality once more yields $\|E_I(P) - E_I(Q)\| \leq BH(P, Q)$. □

The coefficient one is worst-case sharp over this admissible class. For any $B > 0$, take $U = I = \{1, 2\}$, $\mathcal{X} = \mathbb{R}$, and $\mathcal{Z}_{\mathrm{ev}} = \{z_1, z_2\}$ with $z_1 \neq z_2$, where $z_{\mathrm{ev}}(1) = z_1$ and $z_{\mathrm{ev}}(2) = z_2$. Define the fixed aggregator on every subset of $\mathcal{Z}_{\mathrm{ev}}$ by

$$a(\varnothing) = 0, \qquad a(\{z_1\}) = a(\{z_2\}) = B, \qquad a(\{z_1, z_2\}) = -B.$$

For $P = \{\{1\}, \{2\}\}$ and $Q = \{\{1, 2\}\}$, the overlap graph has one erroneous component with $H(P, Q) = 3$, while $E_I(P) = 2B$ and $E_I(Q) = -B$. Hence $\|E_I(P) - E_I(Q)\| = 3B = BH(P, Q)$. This witness establishes worst-case sharpness over the allowed class.

Proposition 12 (Sharp content-aware partition-error bound)**.** *Fix a finite record set I, its canonicalizer, and partitions P, Q of I. For a fixed aggregator a with $\|a(A)\| \leq B$ on every occurring content set,*

$$\|E_I(P) - E_I(Q)\| \leq B\Delta_q(P, Q) \leq BH(P, Q).$$

For each fixed pair P, Q and $B \geq 0$,

$$\sup_{a:\, |a(A)| \leq B} \left| \sum_A (q_P(A) - q_Q(A)) a(A) \right| = B\Delta_q(P, Q),$$

where the supremum is over scalar aggregators on finite nonempty canonical-content sets. An L-Lipschitz downstream margin therefore obeys $|m(E_I(P)) - m(E_I(Q))| \leq LB\Delta_q(P, Q)$, and its nonzero sign is stable if $|m(E_I(P))| > LB\Delta_q(P, Q)$.

Proof. Put $c_A = q_P(A) - q_Q(A)$. The fixed canonicalizer and fixed aggregator give

$$E_I(P) - E_I(Q) = \sum_A c_A a(A).$$

The triangle inequality bounds the norm by $B \sum_A |c_A|$. Identical record blocks cancel between the two partitions. The total numbers of remaining blocks on the two sides sum to $H(P, Q)$; mapping these blocks to their content sets can cancel additional mass, so $\sum_A |c_A| \leq H(P, Q)$. For sharpness, choose the single fixed scalar function $a(A) = B \operatorname{sgn}(c_A)$, with $\operatorname{sgn}(0) = 0$ and zero on absent content sets. It attains $B \sum_A |c_A|$. This is a worst-case choice for the given pair; a particular implemented aggregator can have a smaller distortion. Lipschitz continuity and the strict margin argument prove the final two claims. □

The two discrepancies distinguish record-level and canonical-content-level changes. For example, let $z_1 = z_2 = u$ and $z_3 = z_4 = v$ with $u \neq v$, and compare $P = \{\{1,3\},\{2,4\}\}$ with $Q = \{\{1,4\},\{2,3\}\}$. Their overlap graph has $H(P,Q) = 4$, but both quotient states equal $2\delta_{\{u,v\}}$, so $\Delta_q(P,Q) = 0$ and every fixed set-based aggregator gives identical states. Conversely, the four-distinct-element witness of Proposition 2 has $H = \Delta_q = 4$ and attains distortion $4B$, despite zero change in group count. More generally, Δ_q is a pseudometric on partitions and becomes an ℓ_1 metric on their distinct reachable quotient states.

Corollary 13 (Downstream decision stability)**.** *Let $L \geq 0$ and let $m : \mathcal{X} \to \mathbb{R}$ be L-Lipschitz. Then*

$$|m(E_I(P)) - m(E_I(Q))| \leq LBH(P,Q).$$

For a binary decision boundary at $m = 0$, the nonzero sign is unchanged whenever

$$|m(E_I(P))| > LBH(P,Q).$$

Proof. The first claim follows by applying Lipschitz continuity to Proposition 11. Under the displayed strict condition, the largest admissible change in the margin is smaller than its distance from zero, so the sign remains unchanged. The displayed strict inequality provides a uniform two-sided guarantee: the perturbation remains strictly inside the reference margin. At equality the perturbed margin can reach zero, so a fixed tie convention covers a single reference sign. □

These guarantees compose with an externally specified representation pipeline whose provenance signals, supplied partition, and canonicalized content provide the upstream inputs. Authenticated signatures can carry verified identity without changing the role of opaque membership labels. External caps on the group aggregator instantiate B, and an independently verified Lipschitz bound instantiates L. Behavioral curves test grouping effects; system-level checks provide the numerical stability constants.

C Detailed Experimental Specifications

C.1 Cue and Topology Families

The decision experiment has five prompt families. Marker swaps which value is labeled verified. Graph swaps anonymous record-to-node edges. Provenance uses the same graph with nodes explicitly described as origins. Rule adds a same-origin-counts-once instruction. Full crosses marker correctness and topology correctness. Within a paired family, candidate strings, claims, record order, value-table order, literal offsets, and the full tokenizer multiset are fixed.

C.2 Direct Pipeline Modes

Edge asks whether two specified records share a root. Count asks for the number of distinct roots supporting a deterministically selected abstract value. Winner asks which value has more distinct roots. Opaque oracle supplies the 2:1 count summary after removing the graph. Semantic oracle supplies the same summary alongside the original fact task. End-to-end supplies graph, rule, and fact task together. All modes use direct semantic candidates.

C.3 Repetition Doses

At every dose, two records supporting the factual target arise from two distinct roots. The competitor has one root, whose record is repeated D times. Raw hides root IDs; Prompt preserves all records while exposing the root mapping and deduplication rule. External retains one representative per oracle root and removes root instructions from the final prompt. This stable-representative operator fixes the emitted record count per root. The full-content experiment additionally retains complementary passages within each emitted group. Forward and reverse record orders are both evaluated. The External and Raw $D = 1$ serialized inputs are identical for all 282 item-presentation-order pairs in each checkpoint.

C.4 NATURAL-TEXT SUPPLIED UNITS

The natural-text experiment fixes two domains before model inference. HUMAN EVIDENCE contains 101 PERSPECTRUM claims grouped into 49 dependence components. Each side supplies four different crowd-annotated stance-perspective clusters, matched one-to-one to side-exclusive evidence-unit IDs and distinct normalized and visible texts. Each ID indexes one crowd-annotated stance-perspective evidence unit. WEB CONSENSUS contains 138 ConflictingQA questions in 135 dependence components, with frozen two-model-consensus stance groupings. Each selected eight-record panel has operationally separated registrable domains, canonical URLs, normalized texts, and visible texts across both sides. A WEB unit is a supplied canonical-URL key under these frozen criteria. Publisher identity and ownership enter through explicit authenticated signature coordinates. The two domains are analyzed separately with domain-specific component bootstraps.

Text is normalized with Unicode NFKC, removal of control, format, and private-use characters, and whitespace collapse while preserving ASCII punctuation. Eligible evidence has at least 400 normalized characters. The model-visible field is a 400-character head–tail excerpt comprising the first 200 characters, the literal marker `[...]`, and the last 193. Every Base prompt contains three evidence records; Copy, Prompt, and Four-Unit contain six. Copy and Four-Unit match record count and 2,400 visible evidence characters; The matched six-slot control additionally matches final character, byte, and tokenizer length exactly.

For each claim and attack side, a fixed hash-based selection rule chooses two opposite-side anchors and one dose anchor. BASE contains these three units. COPY repeats the dose anchor four times. PROMPT preserves the same six evidence texts and record count as Copy while showing short opaque aliases and a same-unit-counts-once rule. FOUR-UNIT retains the same anchors and dose anchor, replacing three copies with three other attack-side units. Support and undermine attacks and original and reverse record orders yield 16 trials per claim. Opaque aliases are fixed by canonical role before reversal; reversal preserves every field and changes the record sequence.

Candidates are the direct semantic strings `Yes` and `No`; scoring starts from each tokenizer's native assistant state and termination sequence. The primary outcome is strict selection of the manipulated side, with an exact likelihood tie scored zero. We average two attack sides and two orders within claim before forming contrasts. Each domain uses its own frozen dependence-component bootstrap stream for every contrast and metric; a component joins claims that reuse a supplied key or final visible text. The point estimand remains claim-weighted. Ten thousand replicates use master seed 20260828 with deterministic domain-specific offsets (20260828 for WEB and 20270828 for HUMAN); auxiliary sensitivities use separately frozen offsets, and ordinary item bootstrap provides sensitivity estimates.

Copy–Base tests replication invariance: its 90% interval must lie entirely inside $[-5, +5]$ points. Four-Unit–Base, Four-Unit–Copy, and Copy–Prompt each require a point estimate of at least $+5$ points and a 95% lower bound above zero. These gates are evaluated separately for every checkpoint and domain. Stable collapse of Copy by its hidden unit key reproduces the corresponding Base prompt. Consequently, upstream recovery is numerically identical to Copy–Base and represents the same evidence contrast.

The 5-point operational margin corresponds to approximately one changed choice per 20 constrained decisions. The stricter 10-point marker criterion in the decision experiment was likewise frozen before treatment inference. The two gate types answer different questions: an equivalence band assesses whether an effect is small enough to treat as negligible, whereas a positive-effect gate assesses whether an estimate reaches a stated magnitude with direction supported by its interval. Appendix Table 15 reports a 2.5/5/7.5/10-point sensitivity sweep around the frozen gates. Copy–Base lies outside equivalence in all eight model–domain cells at every margin; model-specific estimates and intervals are primary, and pass counts summarize gate-level magnitudes.

C.5 MODEL CONFIGURATIONS AND GROUPING INTERVENTIONS

Model grid. The natural-text grid is evaluated with the same prompts, candidates, estimands, and bootstrap streams on Qwen3-14B-AWQ and Phi-4. Qwen3-14B uses the publisher's AWQ 4-bit checkpoint; Phi-4 uses dynamic NF4 inference. The Qwen3-8B NF4 rerun is counted as a quantization-control inference configuration under the existing Qwen3-8B checkpoint identity. Each inference configuration completes 3,824 trials, for 11,472 new candidate-choice forward trials.

Controlled answer-plus-one-sentence generation. The generation panel freezes 40 HUMAN and 40 WEB items, one per natural-text dependence component, before inference. For each item it crosses BASE, raw four-copy, and prompt-rule conditions with two attack sides and two record orders, yielding 960 trials per checkpoint. Qwen3-8B runs in bfloat16, Qwen3-14B uses its official AWQ checkpoint, and Phi-4 uses dynamic NF4 inference. Greedy decoding requests a leading `Yes` or `No` followed by one concise sentence, with at most 48 new tokens. All-trial leading-answer selection is the primary endpoint; invalid outputs receive zero attack selection. The valid-output analysis is a frozen sensitivity. The explanatory sentence is retained in the output, and scoring uses deterministic leading-answer parsing.

For each domain and checkpoint, we average the two attack sides and two orders within item, then bootstrap the 40 frozen dependence components 10,000 times. All three formal runs contain 960/960 predictions and zero invalid outputs, for 2,880 controlled-generation trials.

False splits and false merges after fixing supplied keys. This experiment uses the supplied natural-text grouping keys and injects partition error before record emission. Eligibility requires both sides of a claim to contain a frozen document of at least 160 whitespace-delimited words. The resulting HUMAN panel has 66 claims in 32 dependence components; the WEB panel retains 138 questions in 135 components. This filter is model-independent and fixed before treatment inference.

For the one-group panel, the attack material consists of four non-overlapping 40-word windows from one frozen document. For the four-group panel, four 40-word windows come from four supplied operational units. Estimated keys emit one, two, or four records after within-group exact-text deduplication and a fixed per-group content budget. The richer PARTIAL-G1-FULL-CONTENT endpoint concatenates all four unique windows into one 160-word record while keeping estimated group count at one; subtracting PARTIAL-G1 measures the full-window inclusion gain relative to the one-window 40-word endpoint. The main content-fixed contrasts retain all 160 attack words in the one-record condition: PARTIAL-G4 minus PARTIAL-G1-FULL-CONTENT measures false splitting of one supplied document, and DIST-G4 minus DIST-G1-FULL-CONTENT measures the loss from merging four supplied operational units. In each pair, the ordered attack words are identical and record placement implements the grouping intervention.

The primary outcome is the attack-side likelihood share in Equation (7). Secondary endpoints are strict selection, half-tie selection, and likelihood margin. Two attack sides and two orders are averaged within item, and 10,000 bootstraps resample supplied-key or final-text connected components while retaining all member claims. Twelve conditions produce 9,792 trials per checkpoint and 39,168 across Qwen3-8B, Qwen3-4B, Phi-4-mini, and Mistral-7B. Conditioning on the supplied operational keys, the experiment identifies the causal effect of downstream record placement.

Matched six-slot serialization control. The four-record experiment fixes evidence words while jointly varying partition rendering and its deterministic serialization. The original renderer fixes the instruction, query, claim, anchor content, chat adapter, candidates, and scoring suffix. Its four-record arm contains three additional JSON objects, ordinal record IDs, side and evidence headers, and object separators. The serialized difference is 147–153 UTF-8 bytes and 43–63 tokenizer input tokens across the four checkpoints, while the evidence-token difference is bounded by five. Table 6 therefore estimates the total effect of operational record placement under fixed evidence words.

The six-slot control removes these count and length differences. Both arms use six JSON objects with the same R1–R6 identifiers, side labels, headers, separators, instruction, query, claim, anchor content, and candidate format. The same ordered 160-word attack stream is placed in four attack-side slots as either $[160, 0, 0, 0]$ or $[40, 40, 40, 40]$ words; the empty slots and their headers remain model-visible in both arms. A deterministic outcome-blind search adds up to four whitespace characters after the final JSON brace so that each tokenizer-specific pair has exactly the same final character count, UTF-8 byte count, and input-token count while parsing to the same JSON records. Model-visible, parse-inert whitespace supplies the exact length match. The identified treatment is the placement of a fixed word stream across one versus four nonempty content-bearing record slots within a matched six-slot skeleton.

Each checkpoint completes 3,264 choice trials, yielding 13,056 trials and 26,112 candidate forward calls. Two sides and two orders are averaged within item; 10,000 replicates resample the same dependence components used by the four-record experiment. All 16 model–domain–contrast point

estimates have the predicted positive sign, and 12 intervals lie entirely above zero. The matched control identifies a partition-boundary effect beyond extra objects, repeated headers, identifiers, separators, and total tokenizer length.

Table 3: Matched-serialization control. Effects are percentage-point changes in p_{attack} with 95% dependence-component bootstrap intervals. The intervention is matched six-slot serialization; the same ordered evidence word stream is placed across one versus four nonempty content-bearing record slots. Model-visible, parse-inert pair padding equalizes final character, byte, and tokenizer lengths. The estimand is operational content placement across record slots within the matched skeleton.

Model	Domain	One supplied unit: split	Four supplied units: merge loss
Qwen3-8B	WEB (138/135)	+8.27 [+5.73, +10.82]	+8.00 [+5.18, +10.88]
Qwen3-8B	HUMAN (66/32)	+5.92 [+2.39, +8.87]	+11.27 [+7.37, +15.96]
Qwen3-4B	WEB (138/135)	+9.95 [+6.90, +13.04]	+13.33 [+10.29, +16.43]
Qwen3-4B	HUMAN (66/32)	+1.24 [-2.48, +5.69]	+7.47 [+3.57, +12.23]
Phi-4-mini	WEB (138/135)	+1.24 [+0.38, +2.29]	+1.25 [+0.40, +2.39]
Phi-4-mini	HUMAN (66/32)	+0.63 [-0.02, +1.65]	+0.99 [+0.01, +2.44]
Mistral-7B	WEB (138/135)	+3.22 [+1.11, +5.27]	+1.46 [-0.40, +3.34]
Mistral-7B	HUMAN (66/32)	+6.46 [+3.65, +8.84]	+2.33 [-0.65, +5.73]

Controlled GEO proliferation panel. The frozen panel contains 48 fictional product pairs. For each product and attack side, the same-root family uses four exactly 40-word genre-varied records bound to one three-fact campaign brief and one oracle root. The independent-root family uses four exactly 40-word records bound respectively to laboratory, user-panel, endurance, and service-data facts generated by four oracle roots. Every visible claim belongs to its enumerated atomic-fact set. Each attack stream contains exactly 160 ordered words; hidden root, author, and synthetic-domain fields are absent from the prompt. Both structures use the matched six-slot renderer and compare $[160, 0, 0, 0]$ with $[40, 40, 40, 40]$ under exact character, UTF-8 byte, and model-token length matching.

Each checkpoint contributes 768 choices, yielding 3,072 choices and 6,144 candidate forward calls. Every condition–attack-side–order cell contains 48 trials. The primary item-first estimand averages the two attack sides and two orders before taking the contrast; 10,000 fixed-seed bootstrap replicates resample the 48 items. A descriptive mirror decomposition reveals substantial presentation interactions. For Qwen3-8B, the independent-root contrast is $+41.43$ points in original order and -49.92 in reverse order. The main table therefore reports balanced mirrored effects for each checkpoint and preserves both reversals whose pointwise 95% intervals exclude zero.

Grouping baselines and downstream distortion. We evaluate four grouping rules on 192 within-item partition problems containing 768 records. Oracle root identifiers and single-link cosine clustering with MiniLM sentence embeddings at threshold 0.80 recover every partition. Normalized exact hashing and 128-permutation, three-word-shingle MinHash at threshold 0.80 emit all four same-root records as singletons while preserving all four independent roots. On the 96 same-root cases, these lexical methods have pairwise recall and F1 equal to zero, 576 false-split pairs, adjusted Rand index zero, B-cubed F1 0.40, and total $H = 480$. Table 4 reports partition accuracy, and Table 5 maps every predicted partition to the corresponding executed intervention arm. Thresholds were fixed before result inspection.

Table 4: Grouping baselines on the controlled provenance panel. Exact partition rates are over 96 attack-side cases per provenance structure. H is the total partition-error quantity from Proposition 11.

Method	Same-root exact	Independent-root exact	Total H
Oracle root	100%	100%	0
Exact normalized hash	0%	100%	480
MinHash	0%	100%	480
Sentence embedding	100%	100%	0

Table 5: Downstream grouping distortion on same-root campaign cases. Each entry is $|p_{\mathrm{attack}}(\widehat{P}) - p_{\mathrm{attack}}(P^{\star})|$ in percentage points. The predicted partitions coincide exactly with executed matched-intervention arms. Independent-root distortion is zero for all three methods because each preserves the four oracle roots.

Method	Qwen3-8B	Qwen3-4B	Phi-4-mini	Mistral-7B
Exact normalized hash	18.46	9.66	27.44	4.44
MinHash	18.46	9.66	27.44	4.44
Sentence embedding	0.00	0.00	0.00	0.00

D Additional Results

D.1 Content Retention and Partition Decomposition

Table 6: Central causal decomposition of content retention and record partition, in percentage points with 95% dependence-component bootstrap intervals. Full-content gain compares the 160-word one-group record with its 40-word endpoint. The split and merge contrasts hold the ordered 160 attack words fixed; positive values denote influence gained by splitting one supplied group or influence lost by merging four supplied groups. The columns identify complementary-content retention, false-split inflation, and false-merge suppression. Cells report claims/dependence components: HUMAN 66/32 and WEB 138/135.

Checkpoint	Domain	Full-content gain	One supplied unit: split	Four supplied units: merge loss
Qwen3-8B	HUMAN (66/32)	$+13.98$ $[+8.65, +19.82]$	$+26.38$ $[+20.51, +32.36]$	$+25.17$ $[+18.26, +33.11]$
Qwen3-8B	WEB (138/135)	$+14.30$ $[+11.24, +17.51]$	$+25.23$ $[+21.43, +29.19]$	$+24.65$ $[+20.55, +28.77]$
Qwen3-4B	HUMAN (66/32)	$+29.82$ $[+25.39, +34.45]$	$+20.61$ $[+15.38, +25.68]$	$+31.79$ $[+27.05, +38.06]$
Qwen3-4B	WEB (138/135)	$+15.35$ $[+11.80, +18.82]$	$+24.43$ $[+20.99, +28.01]$	$+31.56$ $[+27.46, +35.81]$
Phi-4-mini	HUMAN (66/32)	$+4.96$ $[+2.49, +8.14]$	$+14.46$ $[+9.81, +19.94]$	$+9.13$ $[+6.02, +13.08]$
Phi-4-mini	WEB (138/135)	$+2.91$ $[+1.07, +4.80]$	$+10.27$ $[+7.64, +13.23]$	$+10.10$ $[+7.65, +12.76]$
Mistral-7B	HUMAN (66/32)	$+18.93$ $[+14.27, +24.16]$	$+32.66$ $[+28.59, +37.74]$	$+31.15$ $[+27.18, +35.64]$
Mistral-7B	WEB (138/135)	$+18.47$ $[+14.99, +22.08]$	$+25.21$ $[+21.99, +28.49]$	$+26.69$ $[+23.00, +30.51]$

Table 1 presents the same-root effect as before/after probabilities and relative reductions. Table 6 retains all three original causal contrasts and their percentage-point intervals. Relative reductions in Table 1 are computed from the existing predictions, with both arms resampled together using the same 10,000 dependence-component bootstrap draws as the original contrast. Means remain specific to each model and dataset, and the two attack sides and two order mirrors are averaged within item first.

D.2 Order Decomposition of Controlled Partition Effects

Table 7: Post-hoc order decomposition of the existing controlled GEO predictions. Both families use four content-bearing records minus one grouped record. Entries are percentage points with pointwise 95% bootstrap intervals over 48 item-components after averaging the two attack sides. Interaction is original minus reverse, estimated with paired resampling. Each checkpoint and family is reported separately; the intervals are descriptive. Values smaller than 0.005 points in magnitude round to 0.00.

Model	Roots	Original order	Reverse order	Order interaction
Qwen3-8B	Same root	$+16.83$ $[+12.37, +21.54]$	$+20.09$ $[+15.91, +24.46]$	-3.26 $[-9.91, +3.37]$
Qwen3-8B	Independent	$+41.43$ $[+36.15, +46.22]$	-49.92 $[-49.95, -49.88]$	$+91.36$ $[+86.07, +96.14]$
Qwen3-4B	Same root	-13.97 $[-19.26, -9.39]$	$+33.28$ $[+29.53, +36.85]$	-47.26 $[-53.93, -41.21]$
Qwen3-4B	Independent	$+10.98$ $[+6.43, +15.92]$	-0.19 $[-0.28, -0.12]$	$+11.17$ $[+6.63, +16.10]$
Phi-4-mini	Same root	$+17.95$ $[+14.19, +21.99]$	$+36.93$ $[+33.56, +39.98]$	-18.98 $[-23.42, -14.32]$
Phi-4-mini	Independent	$+19.15$ $[+15.09, +23.29]$	$+12.31$ $[+9.13, +15.79]$	$+6.84$ $[+2.33, +11.35]$
Mistral-7B	Same root	-8.88 $[-14.64, -3.29]$	0.00 $[0.00, 0.00]$	-8.88 $[-14.64, -3.29]$
Mistral-7B	Independent	$+6.03$ $[+3.27, +9.28]$	$+0.01$ $[+0.01, +0.01]$	$+6.03$ $[+3.26, +9.27]$

Table 7 analyzes all 3,072 campaign-panel choices. Each model has 48 items, two root families, two attack sides, two order mirrors, and two partition arms. For each item, family, and order, we subtract the one-group probability from the four-group probability and then average the two attack sides. The original-minus-reverse contrast is formed within the same item. We resample the 48 distinct item-components 10,000 times with seed 20260907, keeping every paired contrast together, and report percentile intervals. These are post-hoc descriptive intervals; the primary balanced estimates and their original intervals remain as reported in Table 2. The decomposition makes the presentation dependence of each checkpoint visible alongside its balanced effect.

D.3 Model and Generation Results

Table 8: Breadth checks at added 14B checkpoints and in controlled generation, in percentage points with 95% dependence-component bootstrap intervals. The generation endpoint is the parsed leading Yes/No answer.

Checkpoint	Domain	Copy–Base	Copy–Prompt
Added 14B checkpoints, complete-candidate likelihood			
Qwen3-14B official AWQ 4-bit	HUMAN	$+51.49[+45.71, +57.43]$	$+4.70[+0.24, +9.23]$
Qwen3-14B official AWQ 4-bit	WEB	$+30.62[+25.90, +35.40]$	$+0.91[-2.14, +3.93]$
Phi-4 14B dynamic NF4	HUMAN	$+51.98[+45.41, +58.82]$	$+9.16[+5.46, +12.99]$
Phi-4 14B dynamic NF4	WEB	$+43.30[+38.13, +48.52]$	$+1.63[-1.48, +4.86]$
Controlled generation, leading-answer endpoint			
Checkpoint	Domain	Leading Copy–Base	
Qwen3-8B	HUMAN	$+20.62[+15.00, +26.25]$	
Qwen3-8B	WEB	$+15.62[+9.38, +22.50]$	
Qwen3-14B-AWQ	HUMAN	$+16.25[+9.38, +23.12]$	
Qwen3-14B-AWQ	WEB	$+16.88[+10.62, +23.12]$	
Phi-4	HUMAN	$+20.00[+13.75, +26.88]$	
Phi-4	WEB	$+20.00[+13.12, +27.50]$	

D.4 Grouping-Key Robustness Curves

The dose curves vary emitted group count together with the fixed 40-word budget per group and show system behavior as an estimated partition moves between one, two, and four groups. Table 6 reports both the one-group full-window inclusion gain and the content-fixed comparisons in which the ordered 160 attack words remain constant while record placement changes.

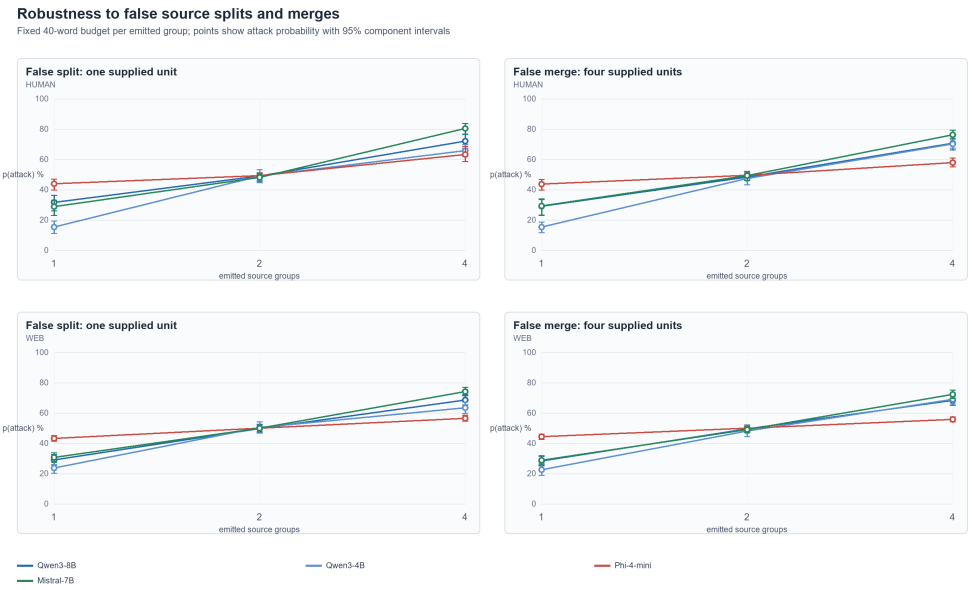


Figure 3: Fixed-budget supplied-key robustness curves. The partial-copy panels split non-overlapping 40-word windows from one frozen document into one, two, or four emitted records. The four-unit panels merge four operationally distinct supplied units into one, two, or four emitted records. The curves jointly vary grouping and the fixed per-group content budget; the content-fixed contrasts in Table 6 provide the primary comparison. HUMAN contains 66 claims in 32 dependence components; WEB contains 138 claims in 135 components. Each domain and checkpoint is plotted separately.

D.5 Multiplicity-Adjusted Bootstrap Sensitivity

Table 9: Bonferroni familywise bootstrap sensitivity. Counts show cells whose two-sided interval has a positive lower bound. Adjusted 95% intervals use percentile $\alpha/(2m)$ tails with $m = 24$ for every Table 6 row and $m = 16$ for the matched six-slot row.

Reported subset	Cells	Ordinary 95%	Familywise 95%
Table 6: full-content gain	8	8/8	7/8
Table 6: content-fixed split/merge	16	16/16	16/16
Table 6: all contrasts	24	24/24	23/24
Matched six-slot control	16	12/16	11/16

We apply Bonferroni-adjusted percentile intervals within two design-distinct result families using 10,000 dependence-component bootstrap replicates regenerated from the frozen seeds. For the 24 Table 6 cells, simultaneous two-sided 95% intervals retain positive lower bounds in 23 cells: all 16 content-fixed split/merge effects and seven of eight full-content gains. The Phi-4-mini/WEB full-content gain is $+2.91$ points with adjusted interval $[-0.02, +5.95]$. Directional familywise lower bounds are positive in all 24 cells. For the 16 matched six-slot effects, 11 simultaneous two-sided intervals retain positive lower bounds. These counts summarize multiplicity sensitivity; the model–domain estimates remain the reported effects.

D.6 Observable Pipeline Stages

Table 10: Equal-domain strict accuracy (%) for observable pipeline stages. Each estimate averages mirrored directions and three presentations within item.

Checkpoint	Edge	Count	Winner	Opaque 2:1	Semantic 2:1	End-to-end
Qwen3-8B	58.85	47.10	51.04	100.00	54.30	49.64
Qwen3-4B	54.26	48.93	48.94	98.93	53.26	49.64
Phi-4-mini	50.00	49.61	51.77	74.85	57.23	49.28

D.7 Decision Contrasts

Across the frozen Marker, M–P, and Rescue gates, all four checkpoints remain below threshold. The provenance-equivalence gate is satisfied by Qwen3-8B, Qwen3-4B, and Mistral; Phi lies outside the equivalence band. The Qwen3-8B tie leaves every gate unchanged under the half-tie sensitivity analysis.

Table 11: Complete-candidate effects on the equal-domain 47-item panel. Likelihood contrasts use 95% item-bootstrap intervals. Accuracy effects are percentage points; provenance uses a 90% interval for the frozen equivalence test and the others use 95% intervals.

Checkpoint	Marker LL	Prov. LL	M–P LL
Qwen3-8B	1.618 $[1.350, 1.876]$	-0.349 $[-0.531, -0.170]$	1.967 $[1.590, 2.321]$
Qwen3-4B	2.175 $[1.343, 3.180]$	-0.243 $[-0.448, -0.045]$	2.418 $[1.553, 3.435]$
Phi-4-mini	1.102 $[0.628, 1.598]$	0.306 $[0.025, 0.602]$	0.796 $[0.248, 1.320]$
Mistral-7B	2.586 $[2.086, 3.115]$	-0.213 $[-0.301, -0.127]$	2.799 $[2.279, 3.347]$

Checkpoint	Marker acc.	Prov. acc.	M–P acc.	Rescue acc.
Qwen3-8B	2.87 $[0.69, 5.74]$	-0.72 $[-2.17, 0.00]$	3.59 $[0.69, 7.25]$	0.00 $[0.00, 0.00]$
Qwen3-4B	2.90 $[0.00, 6.52]$	0.00 $[-1.45, 1.45]$	2.90 $[0.00, 6.52]$	0.00 $[0.00, 0.00]$
Phi-4-mini	3.47 $[-2.90, 9.78]$	2.93 $[-1.36, 7.31]$	0.54 $[-7.94, 8.42]$	0.66 $[-3.62, 5.62]$
Mistral-7B	3.59 $[0.72, 6.52]$	0.00 $[0.00, 0.00]$	3.59 $[0.72, 6.49]$	-1.45 $[-3.62, 0.00]$

D.8 Repetition Contrasts

Phi satisfies the attack, external-recovery, and external-advantage gates. The prompt-recovery pass count is 0/3. Because External equals Raw $D = 1$, external recovery and attack reversal are the same numerical contrast.

The Phi attack remains positive in both domains (30.43 and 23.61 points), all three presentations (38.41, 24.18, and 18.48 points), and both record orders (33.06 and 20.98 points). Across its six presentation-by-order cells, the magnitude spans 4.35 to 48.64 points. Qwen3-4B yields a 4.35-point aggregate change on KNOWN items and five of 47 items.

Table 12: Same-root repetition contrasts in percentage points, equal-domain item means with 95% bootstrap intervals. A pass requires an estimate of at least 5 points and an interval lower bound above zero.

Checkpoint	Raw D1–D8	Prompt–Raw D8
Qwen3-8B	1.42 $[-0.03, 3.20]$	0.35 $[-2.48, 3.56]$
Qwen3-4B	4.35 $[1.09, 8.70]$	0.00 $[-1.81, 1.81]$
Phi-4-mini	27.02 $[21.60, 32.43]$	-7.02 $[-12.94, -0.95]$

Checkpoint	External–Raw D8	External–Prompt D8
Qwen3-8B	1.42 $[-0.03, 3.19]$	1.07 $[-1.45, 3.61]$
Qwen3-4B	4.35 $[1.09, 8.70]$	4.35 $[0.72, 8.70]$
Phi-4-mini	27.02 $[21.63, 32.44]$	34.04 $[27.94, 40.07]$

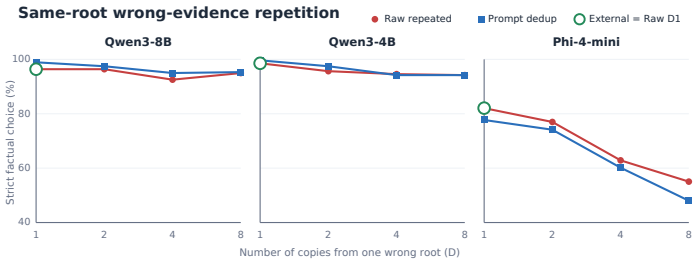


Figure 4: Strict factual choice in the controlled 47-item panel under repeated records from one wrong root. Raw prompts hide shared roots; Prompt prompts show root IDs and a deduplication rule. The green external point is the oracle-folded input and is exactly the Raw $D = 1$ input. The vertical axis starts at 40%; points are equal-domain item means and paired-contrast intervals appear in Table 12.

D.9 NATURAL-EVIDENCE CONTRASTS AND TOKEN SENSITIVITY



Figure 5: Record multiplicity changes complete-candidate decisions in both natural-evidence domains. Copy–Base uses a 90% interval and the shaded ± 5-point invariance band; other contrasts use 95% intervals.

Appendix Table 13 contains all strict selection contrasts. Across eight separately evaluated model–domain cells, replication invariance is 0/8, Four-Unit–Base sensitivity is 8/8, Four-Unit–Copy selectivity is 4/8, and prompt recovery is 3/8. Model-specific effects and intervals remain the inferential unit; these counts summarize the eight frozen decisions.

The near-token-balanced sensitivity subset uses the pre-inference criterion of a mean absolute Copy–Four-Unit input-token difference bounded by 20 over the four attack-side/order cells. The prespecified gates apply to the full analysis.

Table 13: Strict attack-side selection effects in percentage points. HUMAN uses annotated evidence units. WEB uses supplied canonical-URL units under the frozen operational separation criteria; a two-model consensus fixes stance labels. Copy–Base uses a 90% dependence-component bootstrap interval for the frozen ± 5-point invariance gate; the other columns use 95% intervals. Bold cells satisfy their prespecified gate. Every domain and checkpoint is analyzed separately.

Checkpoint	Copy–Base	Four-Unit–Base
HUMAN: annotated supplied evidence units		
Qwen3-8B	$+41.83$ $[+36.73, +47.30]$	$+44.06$ $[+38.11, +50.57]$
Qwen3-4B	$+43.32$ $[+39.24, +47.58]$	$+48.27$ $[+42.93, +53.04]$
Phi-4-mini	$+14.11$ $[+10.82, +17.93]$	$+14.85$ $[+11.25, +19.02]$
Mistral-7B	$+49.01$ $[+44.48, +53.68]$	$+60.89$ $[+54.52, +67.07]$
WEB: operational supplied URL units; frozen stance consensus		
Qwen3-8B	$+31.88$ $[+28.26, +35.51]$	$+40.76$ $[+35.43, +46.17]$
Qwen3-4B	$+34.06$ $[+30.70, +37.41]$	$+42.21$ $[+37.23, +47.08]$
Phi-4-mini	$+9.24$ $[+6.93, +11.69]$	$+9.60$ $[+6.83, +12.59]$
Mistral-7B	$+37.14$ $[+33.33, +40.78]$	$+53.44$ $[+48.33, +58.45]$

Checkpoint	Four-Unit–Copy	Copy–Prompt
HUMAN: annotated supplied evidence units		
Qwen3-8B	$+2.23$ $[-0.74, +5.29]$	-3.47 $[-5.18, -1.72]$
Qwen3-4B	$+4.95$ $[+1.74, +7.80]$	$+5.69$ $[+2.13, +9.31]$
Phi-4-mini	$+0.74$ $[-1.02, +2.39]$	$+1.24$ $[+0.23, +2.67]$
Mistral-7B	$+11.88$ $[+8.43, +15.23]$	$+10.64$ $[+6.99, +14.18]$
WEB: operational supplied URL units; frozen stance consensus		
Qwen3-8B	$+8.88$ $[+4.96, +12.86]$	-2.54 $[-5.40, +0.18]$
Qwen3-4B	$+8.15$ $[+4.35, +12.04]$	-3.99 $[-6.87, -1.09]$
Phi-4-mini	$+0.36$ $[-0.54, +1.27]$	$+0.00$ $[-0.54, +0.54]$
Mistral-7B	$+16.30$ $[+12.41, +20.29]$	$+10.33$ $[+7.06, +13.59]$

Positive values mean more influence from the manipulated side. Certified upstream collapse reproduces BASE byte-for-byte; its recovery is the same paired contrast as Copy–Base.

Table 14: Near-token-balanced sensitivity for Four-Unit–Copy. The subset was frozen from tokenizer counts before model inference. Intervals use the same within-domain dependence-component resampling rule.

Domain	Checkpoint	n	Selectivity [95% CI]
HUMAN	Qwen3-8B	57	$+0.44[-3.19, +4.17]$
	Qwen3-4B	57	$+6.14[+1.96, +9.69]$
	Phi-4-mini	72	$+0.69[-1.72, +2.87]$
	Mistral-7B	50	$+11.00[+5.77, +15.69]$
WEB	Qwen3-8B	79	$+9.49[+4.17, +14.87]$
	Qwen3-4B	79	$+8.86[+3.12, +14.74]$
	Phi-4-mini	89	$-0.28[-1.37, +0.57]$
	Mistral-7B	57	$+14.47[+7.89, +20.61]$

D.10 Threshold Sensitivity

Table 15: Threshold sensitivity. Cells show passes/eligible model–domain cells. Equivalence uses a 90% CI within $[-\delta, +\delta]$; positive effects require an estimate of at least δ and a 95% lower bound above zero. The sweep reports the frozen gates across four operational margins.

Contrast	2.5 pp	5 pp	7.5 pp	10 pp
Marker positive effect	2/4	0/4	0/4	0/4
Provenance equivalence	3/4	3/4	4/4	4/4
Marker–provenance positive effect	2/4	0/4	0/4	0/4
Topology-rescue positive effect	0/4	0/4	0/4	0/4
Copy–Base equivalence	0/8	0/8	0/8	0/8
Four-Unit–Base positive effect	8/8	8/8	8/8	7/8
Four-Unit–Copy positive effect	5/8	4/8	4/8	2/8
Copy–Prompt positive effect	3/8	3/8	2/8	2/8

Copy–Base lies outside equivalence in all eight cells at every tested margin, so the invariance result is stable across the 2.5–10-point sweep. Four-Unit–Base remains positive and material in eight of eight cells through 7.5 points and seven of eight at 10 points. Selectivity changes from 5/8 at 2.5 points to 2/8 at 10 points, while prompt recovery changes from 3/8 to 2/8. Model-specific effects and intervals are primary; gate counts provide an operational digest.

D.11 Controlled-Generation Diagnostics

Table 16: Diagnostics for controlled answer-plus-one-sentence generation. Each model–domain row contains 480 generated trials: 40 items by three conditions, two attack sides, and two orders. Invalid is computed over those 480 trials. Agreement uses the same matched trials; all are generation-valid and candidate-untied. The parsed leading answer is the scored endpoint.

Checkpoint	Domain	Invalid (%)	Choice agreement (%)
Qwen3-8B	HUMAN	0.00	65.62
Qwen3-8B	WEB	0.00	70.83
Qwen3-14B-AWQ	HUMAN	0.00	63.75
Qwen3-14B-AWQ	WEB	0.00	74.17
Phi-4	HUMAN	0.00	65.62
Phi-4	WEB	0.00	77.08

Every formal output has a valid leading answer, making all-trial and valid-output estimates identical. Agreement quantifies consistency between the candidate and generated interfaces, which expose different decision rules while their Copy–Base effects share direction.

E Inference Settings and Evaluation Counts

The original models use bfloat16 inference. Qwen3-14B uses AWQ quantization; Phi-4 and the Qwen3-8B quantization control use dynamic NF4. Candidate scoring is deterministic and generation is greedy. Chat templates, assistant termination sequences, and bootstrap seeds are fixed before treatment inference.

The observable-stage experiment contributes 1,692 trials per model for Qwen3-8B, Qwen3-4B, and Phi-4-mini. The complete-candidate experiment contributes 1,692 trials for each of those models and Mistral-7B. Controlled repetition contributes 2,538 trials for each of the two Qwen models and Phi-4-mini. The natural-text experiment contributes 3,824 trials for each of the four original models; the extended model grid contributes 3,824 trials in each of three inference configurations. Controlled generation contributes 960 trials in each of three models, the content-fixed grouping experiment 9,792 in each of four, the matched six-slot control 3,264 in each of four, and the controlled campaign panel 768 in each of four.

Together these experiments comprise 104,402 evaluation trials across six models. Quantized reruns contribute trials for the same underlying model. Each candidate-choice trial scores two answers, so the matched control's 13,056 choices require 26,112 candidate forward calls and the campaign panel's 3,072 choices require 6,144 calls.

F Generated Answers under Controlled Distractor Proliferation

This extension evaluates whether estimated record groups improve generated answers when a distractor document is presented repeatedly. It reports Mistral-7B-Instruct-v0.3 and the full 14B Phi-4 checkpoint, under exact-copy and overlapping-excerpt additions. Model–condition reporting was selected retrospectively from a broader evaluation; question selection, grouping parameters, and scoring were fixed before answer generation. The intervals below are pointwise estimates for this selected subset, without adjustment for reporting selection. The evaluation supports conclusions about these controlled additions, not general task utility or deployment safety.

Questions and reference groups. We use 120 questions from the 7,405-question HotpotQA distractor development set (Yang et al., 2018), retaining the released questions, answers, ten Wikipedia paragraphs per question, and supporting-sentence annotations. Questions are traversed in a fixed SHA-256 order. Eligibility requires ten distinct, nonempty documents; at least two annotated supporting documents present; at least one nonsupporting paragraph of four or more sentences; at most 3,000 characters per paragraph; and at most 18,000 characters across the ten paragraphs. We select the first 120 eligible questions with no normalized document title shared with an earlier selected question. This procedure reads no model answers or grouping outcomes. The panel contains 92 bridge and 28 comparison questions, 106 span and 14 yes/no answers, and 1,200 distinct document titles. Document identity defines the reference partition. Natural information-source ancestry and cross-document copying are unknown.

Exact-copy and overlapping-excerpt additions. Each paragraph is split at its sentence midpoint into two contiguous excerpts; a one-sentence paragraph remains one record. Among nonsupporting documents with at least four sentences, the target maximizes title-and-body word overlap with the question after fixed stopword removal, with ties broken by a fixed title hash. Supporting documents cannot be targeted. A benchmark distractor is not thereby labelled factually false.

Exact-copy input adds four identical copies of the target's first-half excerpt. Overlapping-excerpt input adds four views of the target: its first w sentences, last w sentences, odd-positioned sentences, and all sentences in reverse order, where $w = \max(\lceil n/2 \rceil + 1, n - 1)$ for a target of n sentences.

These additions rearrange released text without adding new facts: both conditions retain the same distinct visible title–sentence pairs as the original evidence. Every question receives both conditions.

Grouping and retained content. Raw assigns every input record to a singleton. Exact hash groups identical record bodies after Unicode NFKC normalization, case folding, and whitespace normalization. MiniLM uses the fixed all-MiniLM-L6-v2 encoder and single-link cosine clustering at threshold 0.80, unchanged from the earlier grouping experiment. It reads record bodies alone, without titles, questions, answers, supporting-fact annotations, or reference groups. Visible-title grouping uses the normalized document title as an executable metadata comparator. Titles are also visible to the answer model under all other methods.

The grouping encoder has a 256-token input limit; 19 of the 2,816 distinct record bodies exceed it, up to 534 tokens. This limit affects grouping features only. The answer model receives the complete normalized evidence. Each method emits one record per group, retaining every distinct title–sentence pair after NFKC and whitespace normalization. Traversal follows input order and retains the first occurrence of each pair. Original excerpts precede additions. Groups are sorted by their visible ordered contents and assigned consecutive record numbers; a second presentation reverses the group sequence. The question and short-answer instruction are fixed. Answers, supporting-fact annotations, and hidden reference identifiers are absent from the prompt. Content is preserved, while record count and serialized length can change; this extension does not isolate a token-length-matched effect.

Models and execution. Mistral-7B uses bfloat16; Phi-4 uses 4-bit NF4 with double quantization and bfloat16 computation. Both use native user-message templates, greedy decoding with one beam, and at most 64 new tokens. The software is PyTorch 2.12.1 and Transformers 5.14.1 on an RTX 4090 D; Phi-4 additionally uses bitsandbytes 0.50.2. The attention implementation repeats key/value heads before standard scaled dot-product attention and retains the original mask. Inputs span 1,102–5,978 tokens for Mistral and 910–5,232 for Phi-4, below the 8,192-token input budget, with no evidence truncation.

The reported design has $120 \times 2 \times 4 \times 2 = 1{,}920$ aligned conditions per model and 1,322 distinct model inputs per model. Exact input matches reuse a cached deterministic generation. These are computation counts; there are 120 statistical units. All completions are nonempty. The generation limit is reached by 8 distinct Mistral outputs and 14 distinct Phi-4 outputs, appearing in 12 and 20 aligned conditions respectively. These outputs remain in the scoring set and are scored as produced.

Scoring and uncertainty. We apply the released HotpotQA answer scorer to the entire completion, without extracting a preferred answer after generation. Exact match (EM) is primary; answer token F1 is secondary. Empty outputs would score zero. Each question contributes the mean of its two presentation orders. Method effects compare grouping with raw within the same proliferation condition. Pointwise, unadjusted 95% percentile intervals use 10,000 bootstrap resamples of the 120 questions, retaining the paired methods, conditions, and orders together (seed 20260918). Shared deterministic outputs are not additional independent observations. Tables 17 and 18 include every grouping comparator in both reported conditions.

Table 17: Generated-answer correctness under controlled distractor proliferation. EM and token F1 are percentages with pointwise 95% question-bootstrap intervals. Each cell averages 120 questions and two orders; full completions are scored.

Model	Grouping	EM (%)	F1 (%)
Exact copies			
Mistral-7B	Raw	22.50 [16.25, 29.17]	28.35 [22.11, 34.81]
Mistral-7B	Exact hash	26.25 [19.58, 33.33]	33.34 [26.65, 40.27]
Mistral-7B	MiniLM	26.67 [20.00, 33.75]	33.24 [26.41, 40.30]
Mistral-7B	Visible title	26.67 [20.00, 33.75]	33.76 [27.08, 40.79]
Phi-4	Raw	30.42 [23.33, 37.92]	44.79 [38.44, 51.48]
Phi-4	Exact hash	32.50 [25.42, 40.00]	47.51 [40.97, 54.38]
Phi-4	MiniLM	31.25 [24.17, 38.75]	45.29 [38.77, 52.11]
Phi-4	Visible title	34.17 [26.67, 42.08]	47.82 [40.99, 54.83]
Overlapping excerpts			
Mistral-7B	Raw	20.00 [14.17, 26.25]	25.93 [19.90, 32.41]
Mistral-7B	Exact hash	20.42 [14.58, 26.67]	26.19 [20.18, 32.63]
Mistral-7B	MiniLM	25.42 [18.75, 32.50]	31.25 [24.64, 38.22]
Mistral-7B	Visible title	26.67 [20.00, 33.75]	33.76 [27.08, 40.79]
Phi-4	Raw	28.75 [22.08, 35.83]	43.82 [37.49, 50.48]
Phi-4	Exact hash	28.75 [22.08, 35.83]	43.82 [37.49, 50.48]
Phi-4	MiniLM	34.58 [27.08, 42.50]	48.38 [41.48, 55.41]
Phi-4	Visible title	34.17 [26.67, 42.08]	47.82 [40.99, 54.83]

Table 18: Paired changes relative to raw input under the same proliferation condition, in percentage points. Intervals are pointwise, unadjusted 95% question-bootstrap intervals. All three grouping comparators are reported for both models and conditions.

Model	Grouping	Δ EM (pp)	Δ F1 (pp)
Exact copies			
Mistral-7B	Exact hash	+3.75 [0.83, 7.08]	+4.99 [2.26, 8.20]
Mistral-7B	MiniLM	+4.17 [0.83, 7.50]	+4.88 [1.79, 8.26]
Mistral-7B	Visible title	+4.17 [0.00, 8.75]	+5.41 [1.50, 9.53]
Phi-4	Exact hash	+2.08 [−1.25, 5.42]	+2.72 [−0.47, 6.12]
Phi-4	MiniLM	+0.83 [−2.92, 5.00]	+0.50 [−3.41, 4.52]
Phi-4	Visible title	+3.75 [0.00, 7.50]	+3.03 [−0.28, 6.50]
Overlapping excerpts			
Mistral-7B	Exact hash	+0.42 [0.00, 1.25]	+0.25 [0.00, 0.75]
Mistral-7B	MiniLM	+5.42 [2.08, 9.17]	+5.31 [1.68, 9.33]
Mistral-7B	Visible title	+6.67 [2.50, 11.25]	+7.83 [3.46, 12.41]
Phi-4	Exact hash	+0.00 [0.00, 0.00]	+0.00 [0.00, 0.00]
Phi-4	MiniLM	+5.83 [1.67, 10.00]	+4.57 [0.59, 8.72]
Phi-4	Visible title	+5.42 [0.83, 10.00]	+4.00 [0.14, 8.14]