RESOURCE ATTRIBUTION IN RAG: SELECTION, VERIFICATION, AND UTILIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop a resource attribution framework for recovery after limited fact verification in retrieval-augmented generation. Grounded in statistical decision theory and information-channel comparisons, the framework separates ideal information advantages from rendering loss and a fixed generator's utilization gap. It distinguishes privileged selection of existing evidence from acquiring verification messages, and derives matched controls for attributing observed recovery. A 200-group study on HotpotQA and MuSiQue instantiates these controls across content availability, claim status, and answer selection. Same-text verification under competition raises reference-alias EM by 28.5 and 34.0 points for two Qwen generators, with a smaller, dataset-dependent interaction for Mistral. Retrieved-document controls reveal a different attribution: matches rise from 104/200 to 135/200 when an already-visible oracle-selected quotation is re-exposed as ordinary evidence, then to 137/200 when it is marked verified. The incremental EM effect is 1.0 point $[-2.0, 4.0]$, with secondary benefits in F1 and replacement matching. A frozen answer-blind selector reaches 98/200, failing to retain the oracle gain. These results demonstrate how the framework separates protocol-specific recovery effects and limits their extrapolation, while keeping reference recovery distinct from appropriate abstention. To facilitate further research in this area, we release our code anonymously at `https://anonymous.4open.science/r/fsws62dwas212fdf345`.

1 INTRODUCTION

Retrieved knowledge extends a generator's parametric memory (Lewis et al., 2020), but also admits knowledge-database poisoning (Zou et al., 2025). Correcting a poisoned answer leaves a practical question: which dependent questions can now be answered reliably? Verifying a work's author does not also establish that author's birthplace. The successor relation may be missing, present among competing claims, or verified. A recovery gain can therefore reflect added evidence, privileged selection of existing evidence, changed presentation, or changed answer coverage. These possibilities require different comparisons (Figure 1).

We develop a *resource attribution framework* for recovery beyond an initially verified fact. Grounded in statistical decision theory and information-channel comparisons, the framework models the selector's access, the resources it discloses, the actual rendered input, and the generator's decision. Its risk decomposition separates ideal information advantages from rendering loss and generator excess risk. This distinction leads to a concrete evaluation method: retain the evidence interface, expose the same selected span in both arms, and only then vary its verification display.

Our contribution has three parts. First, the framework formalizes why a span already present in context may still convey privileged information through its selection, and why a fixed positive verification marker must be distinguished from acquiring a variable verifier message. Second, we instantiate the framework with content, selection, and matched-display controls. On retrieved documents, matches rise from 104/200 to 135/200 when an oracle-selected quotation is re-exposed as ordinary evidence, then to 137/200 when its status changes. A frozen answer-blind selector instead obtains 98/200, delimiting the oracle result's operational reach. Third, paired response paths and coverage-aware comparisons separate reference restoration from answer suppression. The study combines

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.

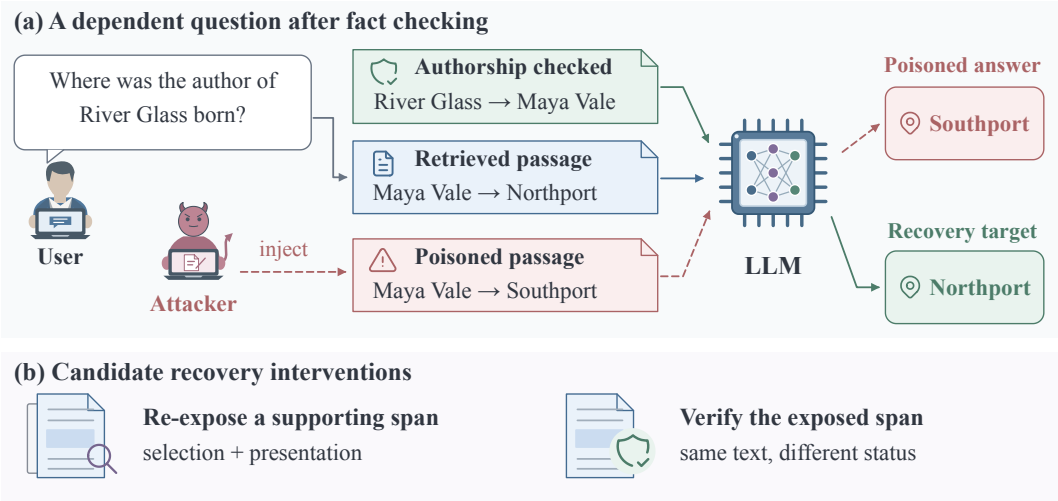

Figure 1: Motivating scenario. Checking authorship leaves a dependent birthplace question exposed to competing retrieved claims. Re-exposing a supporting span and verifying that same span are different candidate interventions. Names and answer paths are illustrative; colors are not truth labels supplied to the generator. Devil adapted from OpenMoji (Mariella Steeb, CC BY-SA 4.0).

200 groups from HotpotQA and MuSiQue, two full-protocol generators, a third core replication, and earlier graph pilots; replayed conditions do not add independent groups.

2 Related Work

Poisoning defenses and recovery. PoisonedRAG (Zou et al., 2025) demonstrates targeted answer manipulation through injected knowledge-base text. RobustRAG (Xiang et al., 2024) isolates retrieved passages and aggregates responses, with certified bounds under its stated corruption model. TrustRAG (Zhou et al., 2025) combines cluster filtering with model self-assessment. These are defense mechanisms; our framework evaluates what resource changes account for a recovery gain. RAGForensics (Zhang et al., 2025) already packages verified triples with support text and priority for post-cleaning correction, while HoG-GRAG (Noughabi et al., 2026) addresses support completion and final-hop disambiguation. We distinguish ordinary successor evidence, privileged localization, and claim-status display within these recovery questions.

Context use and answer selection. Context-faithful prompting (Zhou et al., 2023) studies conflict and abstention through prompting rather than added training. Sufficient Context (Joren et al., 2025) separates sufficiency from utilization; FaithEval (Ming et al., 2025) distinguishes faithfulness under unanswerable, conflicting, and counterfactual contexts. We build on these distinctions while separating external messages from fixed positive markers. Our local gate adapts the released sufficient-context prompt, not its original Gemini implementation or learned pipeline. Selective classification (El-Yaniv & Wiener, 2010) motivates reporting quality with coverage, without transferring its guarantees to our gates. ReliabilityRAG (Shen et al., 2025) and ReliableRAG (Jiang et al., 2026) estimate source or reasoning reliability; this is distinct from our supplied oracle annotations and fixed-generator display effects.

Selection and attribution. RECOMP (Xu et al., 2023) compares learned compression with an outcome-optimized oracle, whereas Lost in the Middle (Liu et al., 2023) controls relevant-document position. Selection and position effects are therefore established antecedents. We formalize the selector's access and actual observation, retain the context, and freeze the exposed span before varying its status. Building on information-channel comparisons (Blackwell, 1953; Cover & Thomas, 2006), this construction specifies attributable intervention effects while distinguishing them from unestimated latent information advantages. Appendix F compares the relevant interfaces and outcome objectives.

Table 1: The six confirmation conditions. Each row is crossed with no competitor (0) or one competitor (1). B and C have byte-identical claim text; only the successor verification flag changes.

Condition	Successor content	Verified claims
A0 / A1	Correct successor absent	First hop only
B0 / B1	Correct ordinary successor present	First hop only
C0 / C1	Identical to B0 / B1	First hop and correct successor

3 Task and Resource Interventions

Each group contains a natural two-hop question, bridge entity, first-hop fact, and successor fact needed for its final answer. A verification record marks a qualified atomic claim, rather than all text from its source. All conditions supply the same verified first hop. The output is a short answer or `NA`. A generator may still use parametric knowledge, so absence of visible successor evidence does not logically force abstention.

Threat model. We simulate targeted factual poisoning through fixed successor-value substitutions, prepared without feedback from test answers. The core packets contain at most one competing claim. Questions, generator instructions, and the verification channel are held outside attacker control. The retrieval corpus contains one poison note per group, so several groups' notes may be co-retrieved. This evaluates limited knowledge poisoning under an authentic verification channel; adaptive attacks and forged verification signals require separate evaluation.

Four opaque claim slots have a fixed order within a group. Original neutral text fills unused slots. Slot count is matched, but token count is not. Competing claims replace the successor target with another value, preserving the relation and qualifiers. They disagree with the benchmark reference but need not be logically exclusive for every real-world multivalued relation. Verification status is explicitly restricted to the marked claim. A $\rightarrow$ B estimates the effect of making ordinary content visible; B $\rightarrow$ C estimates the effect of supplying a verification flag for already-visible content, conditional on the frozen prompt and selected groups. C requires an additional external verification and is not an equal-information algorithm comparison.

4 A Resource Attribution Framework

Building on statistical decision theory and comparison of information channels (Blackwell, 1953; Cover & Thomas, 2006), we develop a framework for attributing RAG recovery to selection, verification, and utilization. Its purpose is to specify which resource changes a comparison measures. The construction adds a selector-access boundary, an explicit message/display distinction, and an observation-aware decomposition from which we derive matched controls.

Resource and observation interfaces. Let Z contain the question, full visible context, existing verification, and shared background; Y is the target. A selection S includes the exposed span, location, and source binding. An answer-blind policy samples S_b from Z with randomness independent of Y given Z. An oracle policy may additionally use annotations O outside Z, yielding S_o. Blindness covers candidate construction, parameters, caches, and tuning, not only inference-time fields. A verification process can disclose a message M; a renderer produces the actual observation X_j for a fixed generator g. Message acquisition and marker display are distinct interventions.

For the retrieval arms, the disclosed resources are

$$W_d = Z, \quad W_b = (Z, S_b), \quad W_u = (Z, S_o), \quad W_v = (Z, S_o, M).$$

The renderer uses only its disclosed resources and target-independent randomness: $Y \perp X_j \mid W_j$. All arms share a population law, target, protocol knowledge, finite action set, and fixed bounded loss ℓ. For any observation W, define

$$R^*(W) = \mathbb{E}\Big[\min_a \mathbb{E}[\ell(Y,a) \mid W]\Big], \qquad R_g(j) = \mathbb{E}[\ell(Y, g(X_j))]. \tag{1}$$

Generator randomness is independent of the target given its input. We separate rendering loss $C_j = R^*(X_j) - R^*(W_j) \geq 0$ from generator excess risk $E_j = R_g(j) - R^*(X_j) \geq 0$, and write $H_j =$

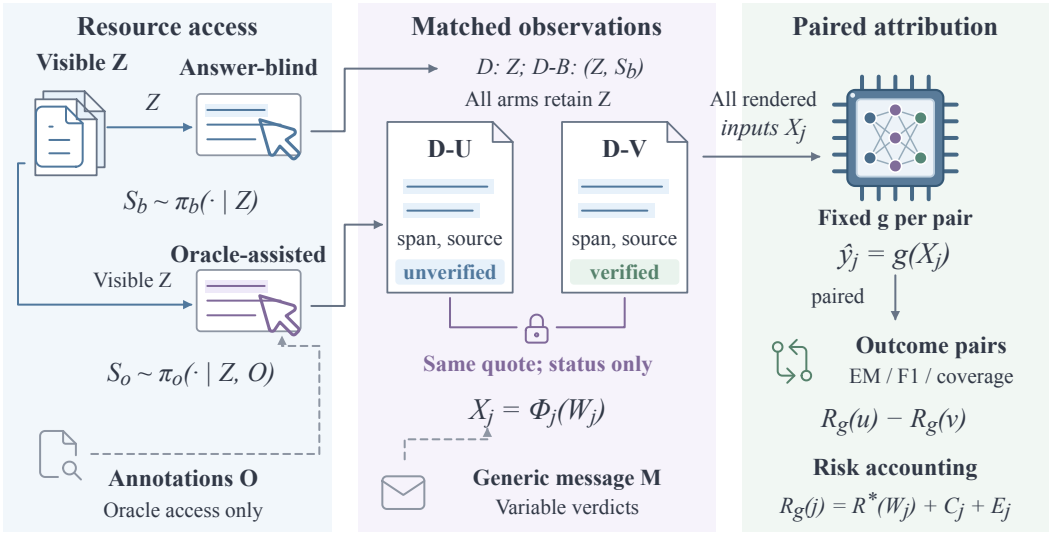

Figure 2: Resource attribution framework. All arms retain Z; only oracle selection accesses O. D-U/D-V share the displayed span and source binding, changing its status. Weights and decoding protocol are fixed within each pair. Dashed M denotes a general message interface, not varying verdicts tested here. Risk accounting is theoretical; the experiments estimate paired outcome effects, not latent information advantages.

$C_j + E_j$. If the disclosed resources are recoverable from the actual observation, $C_j = 0$. Retaining document text alone need not preserve its source binding or selection cue.

Proposition 1 (resource attribution). Define the task-specific ideal advantages

$$A_S = R^*(Z) - R^*(Z, S_o), \qquad A_M = R^*(Z, S_o) - R^*(Z, S_o, M).$$

Both are nonnegative. Under the above interfaces,

$$\begin{aligned} R_g(d) - R_g(b) &= H_d - H_b, \\ R_g(d) - R_g(u) &= A_S + H_d - H_u, \\ R_g(u) - R_g(v) &= A_M + H_u - H_v. \end{aligned} \tag{2}$$

Derivation. Conditional independence gives $R^*(Z, S_b) = R^*(Z)$. An observer can ignore additional selection or message information, proving $A_S, A_M \geq 0$. Substituting $R_g(j) = R^*(W_j) + H_j$ yields the identities. Appendix A proves the rendering inequalities and gives counterexamples.

The decomposition is an operational specialization of classical decision principles. It shows why an oracle-versus-blind performance gap cannot alone measure the oracle's information advantage: it also contains $H_b - H_u$. Even a span already present in Z may disclose information through the oracle's choice. Conversely, an answer-blind repetition can help a fixed generator without adding information. These latent terms are not estimated from our table of unique questions.

Proposition 2 (fixed-marker display equivalence). Suppose U and V expose the same selected span under the same known selector law, and their actual inputs differ only by a reversible deterministic rewrite of a fixed positive marker, without truncation. Then

$$R^*(X_u) = R^*(X_v), \qquad R_g(u) - R_g(v) = E_u - E_v. \tag{3}$$

Each observation can simulate the other, so their ideal risks coincide. Thus a fixed positive-marker contrast measures generator utilization under that experimental law; it is not an estimate of the information supplied by a deployed verifier with variable verdicts. The result applies to our identifiable appended quotation slot. It does not automatically apply when identifying which claim to mark itself requires additional oracle information.

Table 2: The framework's attribution report. Pairwise effects concern a fixed generator and outcome. Access, full rendered inputs, seed, and the paired interval accompany each report.

Contrast	Controlled intervention	Licensed interpretation
D → D-B	Append a frozen answer-blind selection; retain context	Total effect of this blind selection and presentation policy
D → D-U	Append an oracle-localized span already in context	Joint privileged localization and presentation effect
D-U → D-V	Fix span, position, source binding, and text; change its status	Positive-marker display effect; utilization difference under Proposition 2
D-B → D-U	Replace one fixed selection policy by the other	Policy/interface gap, including selection and utilization differences

Example: location versus verdict. Let the visible context contain both candidate values 0 and 1, with Y a uniform bit. If an oracle first selects $S_o = Y$, that location already reveals Y to an ideal observer who knows the selection rule; adding a constant positive mark supplies no further information. If the selected location is instead fixed at 0 and a verifier returns $M = \mathbb{1}\{Y = 0\}$, its varying verdict distinguishes the two states. Both interfaces can be useful to an LLM, but they test different resources.

Reusable attribution report. Figure 2 turns these interfaces into a comparison procedure. We record the allowed selector inputs, retain the exact rendered observations, freeze the paired intervention and outcome, and report its average effect with a paired interval and decision-state counts. Table 2 assigns a licensed interpretation to each contrast. These are observable intervention effects; the latent A_S, A_M, and utilization terms are not numerically identified by the EM table. A → B and B → C provide complementary content-availability and claim-status controls. Differences add along a fixed path, without defining an order-independent allocation. Reference EM supplies a reference-agreement loss, not a label for safety or justified abstention.

5 Experimental Design

Groups and models. We select 100 groups per dataset from HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022), excluding 148 previously used query IDs and 693 source titles. All 400 supporting documents are distinct across these groups. Fixed hash order, automatic source checks, exact-quotation checks, and one bounded replacement-type repair precede test answering; selection does not use evaluation predictions or attack success. The construction model, standard Qwen3.8-27B Q6_K, is also evaluated, so curation bias is possible. There are no new independent human labels. Appendix B records construction, fallback rules, and full settings.

Qwen3-32B-AWQ (Qwen Team, 2025) and the pinned standard Qwen3.8 checkpoint (Unsloth, 2026) run the full protocol: temperature 0, seed 20260916, thinking disabled, and 8192-token context budgets. Model size, quantization, and backend differences are not isolated. After observing the first Qwen generator's core results, we add Mistral-7B-Instruct-v0.3 BF16 (Mistral AI, 2024) on the same six core conditions, with unchanged inputs and no tuning. This bounded addition is not retrospectively part of the original two-model plan.

Decision and retrieval controls. All six conditions receive direct answers. In B0/B1, the same candidate is also passed to ordinary review or the released Sufficient Context autorater prompt, which keeps the candidate or abstains. The gate sees no reference answer or successor oracle. Answer interfaces use a common JSON schema and 128-token limit. Initial gate judgments have 768 tokens; a separately declared policy completes only truncated judgments once at 4096 tokens, preserving candidates and completed judgments. Both budgets and cumulative costs are reported.

Claim order is balanced within dataset; 40 predetermined groups additionally swap relevant slots in A1/B1/C1. Each dataset also has a 400-document corpus of 200 supports, 100 replacement notes, and 100 fillers. MiniLM-L6-v2 retrieves four documents using the question alone or the question plus the verified bridge. Retrieved documents remain unverified; the same first-hop verification is supplied in both arms. Retrieval encodes at most 512 tokens per document, and generation sees

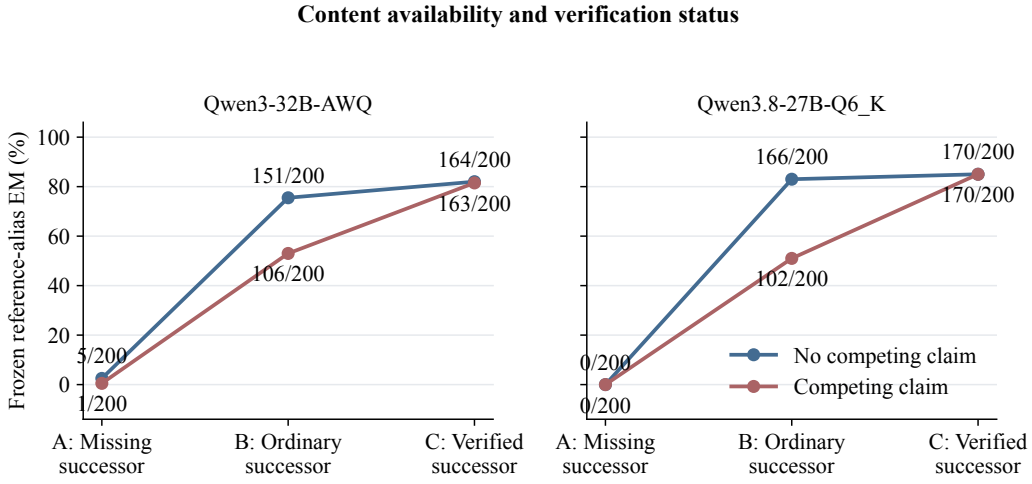

Figure 3: Reference-alias EM over the same 200 fact groups in each condition. A $\rightarrow$ B adds ordinary successor content; B $\rightarrow$ C holds claim text fixed and adds successor verification. All denominators retain interface failures. Curves compare information resources, not equal-resource defenses.

prefixes of at most 1200 characters. This is a flat-retrieval control, not a matched graph-versus-flat comparison.

Matched quotation and answer-blind controls. Two sequential follow-ups use the same 200 bridge-retrieved contexts and Qwen3-32B. D reuses native contexts and outputs. D-U appends an oracle-selected supporting quotation only if it is already fully visible; D-V changes only that appended quotation's status. For nine missing quotations, both inputs and outputs equal D. The 400 views require 382 new responses. D-B then evaluates one frozen RECOMP NQ selector, using only the question, visible first-hop verification, and all four retrieved prefixes. We retain full context and append its top-ranked span as UNVERIFIED. A 128-token span cap leaves all old oracle quotations unchanged; span lengths are not individually matched. All 200 D-B answers are newly generated. The selector, candidate rules, and five-point practical EM noninferiority margin relative to D-U are fixed before these answers. These are component adaptations and resource controls, not end-to-end rankings of the cited methods. Appendices C and D detail the two follow-ups.

Outcomes and accounting. Frozen reference-alias EM measures reference recovery, supplemented by token F1, strict `NA`, interface failure, and designated replacement matching. Nonmatch does not establish factual error. Coverage is valid non-`NA` frequency, and selective quality is reference matching among those answers. For a gate retaining k of N valid candidates with M matches, random retention gives kM/N expected matches at the same coverage, without additional calls. Appropriate abstention remains a distinct, unlabelled objective.

Within-group differences use 2000 bootstrap resamples stratified by dataset; intervals are exploratory without multiplicity correction. The original two-model protocol has 5040 logical records, the Mistral check 1200, and the quotation and blind-selection follow-ups 400 and 200. Shared candidates, retries, and gate completions are separately accounted for. All follow-ups reuse the original groups; the independent sample size remains 200.

6 Confirmation Results

6.1 Content and verification have different marginal value

Without a competing claim, adding ordinary successor content changes Qwen3/Qwen3.8 matches from $5/0$ to $151/166$, out of 200 (Figure 3). Paired gains are 73.0 points [67.0, 79.0] and 83.0 points [77.5, 88.0]. Adding a competing claim reduces these ordinary-content counts to $106/102$. With the

correct successor absent, the competing claim instead produces 147/176 designated replacement-value matches, showing that first-hop verification alone does not prevent adoption of a benchmark-inconsistent successor value.

The same-text B1 $\rightarrow$ C1 intervention raises matches from $106 \rightarrow 163$ and $102 \rightarrow 170$: gains of 28.5 points [22.0, 35.0] and 34.0 points [27.5, 40.5]. Without competition, the corresponding verification effects are only 6.5 and 2.0 points. The verification-by-competition interaction is therefore 22.0 points [15.5, 28.5] and 32.0 points [24.5, 39.0]. Its direction agrees across datasets: 18/26 points on HotpotQA/MuSiQue for Qwen3 and 25/39 for Qwen3.8. Token F1 also increases under competing-claim verification, by 26.0 and 32.7 points (Appendix B).

These gains do not arise solely from a more permissive interpretation of answer wording. B1 $\rightarrow$ C1 produces 58 gains and one loss for Qwen3, and 70 gains and two losses for Qwen3.8. Of the gains, 34 and 32 replace an *exact designated replacement answer* with a reference match. The intervention identifies and marks the disputed successor as externally verified. It establishes a conditional claim-status benefit beyond content completion; the framework does not identify that performance gain with a measured amount of information.

Restoration versus other new matches. A post-hoc B0/B1/C1 decomposition makes the reference-recovery target explicit. Competition removes 48 original matches for Qwen3, 68 for Qwen3.8, and 20 for Mistral. C1 restores 42, 65, and 13 of these, respectively; the remaining six, three, and seven remain nonmatches. The corresponding numbers of C1 gains from groups already nonmatching in B0 and B1 are 16, five, and 29. Thus the same B1-to-C1 gain count can contain different mixtures of restoration and general answer improvement. These are descriptive reference-match paths, not semantic error or appropriate-abstention labels (Table 19).

Cross-family replication bounds the interaction claim. Mistral's A0/A1/B0/B1/C0/C1 match counts are $5/3/71/70/97/110$, out of 200. Ordinary-content gains remain positive (A0 $\rightarrow$ B0: 33.0 points [26.5, 39.5]), as do same-text verification gains under competition (B1 $\rightarrow$ C1: 20.0 points [14.0, 26.0]). However, its verification-by-competition interaction is smaller: 7.0 points [0.0, 13.5], comprising 13.0 on HotpotQA and 1.0 on MuSiQue; the latter interval spans -9.0 to 10.0. B0 $\rightarrow$ B1 has an almost flat net change ($71 \rightarrow 70$), but this hides 19 gains and 20 losses. Thus content and verification benefits extend to this generator, while the strong competition interaction observed for the two Qwen generators should not be generalized to every model. Full results appear in Tables 14–15.

6.2 Retrieved quotations change the attribution of recovery

Table 3: Retrieval-based attribution on the same 200 groups (Qwen3-32B).

Input	EM count	F1%	NA	Repl.
D	104	64.5	3	55
D-B	98	61.3	3	56
D-U	135	79.6	2	24
D-V	137	81.8	2	16

D: native retrieved prefixes. D-B: append a frozen answer-blind selection. D-U: additionally display an oracle-selected quotation already visible in a retrieved prefix, as ordinary evidence. D-V: identical to D-U except that quotation is verified. For D-U/D-V, nine missing-quotation groups retain the identical native input and output. Repl. is designated replacement matching, not a comprehensive factual-error measure.

The original bridge-augmented retrieval run exposes the exact successor quotation for 191/200 groups and retrieves the target group's poison note for 190/200. Source presence does not by itself establish a logically unresolvable conflict. We reuse these document prefixes and the native Qwen3 answers (D), then compare an explicitly displayed but unverified quotation (D-U) with the identical quotation verified (D-V). No missing quotation is inserted. Both added-quotation conditions receive the same oracle localization, so their difference isolates the incremental status flag after this presentation step.

Reference matches change $104 \rightarrow 135 \rightarrow 137$ (Table 3). D $\rightarrow$ D-U gains 15.5 points [10.0, 20.5], comprising 34 newly matched answers and three losses. D-U $\rightarrow$ D-V gains only 1.0 point [-2.0, 4.0], comprising six gains and four losses. Its dataset effects are -2.0 points on HotpotQA and

$+4.0$ on MuSiQue; both EM intervals include zero. The visible-quotation subset gives a similar incremental estimate, 1.05 points $[-2.09, 4.19]$.

The status flag still increases token F1 by 2.2 points [0.1, 4.6], and designated replacement matches fall $24 \rightarrow 16$. Thus the EM interval does not establish absence of every verification benefit. Along this fixed intervention path, however, the 16.5-point end-to-end EM gain from D to D-V cannot be assigned to the verification flag alone: 15.5 points are already obtained by the shared localization/repetition step. This is a concrete change in the attribution of a reported recovery gain. The localization is oracle-assisted and not a free or autonomous alternative to verification.

Where the changed answers go. The saved D-U $\rightarrow$ D-V transitions explain the secondary gains: of eight departing replacement matches, five become reference matches and three become other nonmatches; no new replacement match appears. One additional other-to-reference gain and four reference-to-other losses yield the net two-match increase. Thus the status display changes answer composition beyond its small net EM effect, without establishing that every replacement departure is safe (Appendix E).

The large flag effects in the controlled atomic-claim packets therefore do not establish a large incremental EM effect for this retrieved-document protocol. These interfaces also differ in verification unit and competing material, so their difference is not an isolated causal effect of document format. The comparisons identify path-specific intervention effects. Proposition 2 further distinguishes the reversible status display here from acquiring a variable verification message.

An executable selection policy does not retain the oracle gain. The frozen answer-blind D-B condition yields 98/200 matches: an EM difference of -18.5 points $[-24.0, -13.0]$ from D-U, exceeding the prespecified five-point loss margin. Relative to D, the difference is -3.0 $[-7.0, 1.0]$, with five gains and 11 losses; we do not establish a general degradation or equivalence. The selector chooses the target first-hop document in 159 groups and the successor document in 28. This post-selection source diagnostic suggests a mismatch with successor-evidence needs but does not isolate its cause. The result bounds this fixed configuration, not all answer-blind selection. As Equation 2 makes explicit, its gap to the oracle is not a direct estimate of privileged information value.

Of the 39 D-U-to-D-B match losses, 30 occur among the 159 groups whose selected span comes from the first-hop source; the other source groups contribute four successor-source, four poison-source, and one other-source loss. Within the first-hop-source group, median lexical coverage of first-hop tokens is 0.80. Because the query includes that first-hop text, these observations motivate sensitivity to query construction, but cannot distinguish it from NQ transfer or candidate granularity. We retain this one configuration as an applicability check rather than infer that blind selection generally fails.

6.3 Reference recovery and answer selection

Ordinary B1 review changes Qwen3/Qwen3.8 reference matches $106 \rightarrow 98$ and $102 \rightarrow 53$, with no new matches; coverage changes $95.0\% \rightarrow 87.5\%$ and $81.5\% \rightarrow 40.0\%$. The completed sufficient-context gate retains 18/15 answers with 4/5 matches, below the same-coverage random-retention expectations of 10.04/9.39. These are reference-selection outcomes: a gate may appropriately reject an insufficiently supported reference-matching answer. They do not establish a general reliability ranking. Appendix G retains full operating points, the original truncations, the declared completion policy, and cumulative costs.

6.4 Retrieval coverage and ordering qualify recovery

Bridge-augmented dense retrieval increases successor-document recall from $121/200$ to $192/200$; the exact supporting quotation is visible in $120 \rightarrow 191$ retrieved prefixes. Reference matches rise $70 \rightarrow 104$ for Qwen3 and $73 \rightarrow 102$ for Qwen3.8, paired improvements of 17.0 points [10.5, 23.5] and 14.5 points [8.0, 20.5]. At the same time, retrieval of at least one poison note increases $159 \rightarrow 194$ queries, and designated replacement matching remains $53 \rightarrow 55$ in both generators. Improved retrieval exposes more useful content alongside more poisoned content; high successor recall alone does not establish answer recovery. This comparison does not identify which co-retrieved document caused an individual nonmatch.

In the predetermined 40-group order replay, swapping the relevant claim slots changes B1 EM status on 16 groups for each generator: six gains/ten losses for Qwen3 and five gains/eleven losses for Qwen3.8. Under C1, only four and one EM statuses change, respectively. The net B1 differences have wide intervals (Appendix Table 10); this limited replay supports reporting paired instability rather than claiming general order robustness.

7 Discussion and Limitations

The attribution report links each measured effect to its controlled intervention (Table 2). Oracle selection may disclose a useful location even when it adds no new document text; fixed-generator improvements need not track the ideal information ordering. In the retrieved-quotation experiment, the status contrast follows a shared oracle selection and reversible positive-marker display. It measures utilization of that display under the specified regime, rather than a deployed verifier's information advantage. The frozen answer-blind selector does not retain the oracle gain, adding a concrete applicability boundary without establishing that all answer-blind policies fail.

Our flags are accurate oracle annotations under a protected channel. Real verification may produce negative, uncertain, or wrong messages, and its acquisition cost and accuracy remain untested. The ideal quantities in Equation 2 are explanatory population terms, not estimated mutual information or recoverable components of the EM table. Differences between short claim packets and retrieved documents also change competing material and verification unit; they do not isolate document format. The framework specifies which controls to compare without eliminating these scope conditions.

Reference recovery and justified abstention differ. A gate may appropriately reject an unsupported reference-matching candidate; coverage loss alone does not establish unsafe behavior. Automated curation, selected two-hop groups, repeated fillers, and reference-inconsistent replacements simplify real corpora. EM can miss valid answers and does not certify grounding. The construction model overlaps one generator; Mistral covers only core conditions, and both retrieved-quotation follow-ups use one generator and historical native responses. Numerical run-to-run variation, broad order robustness, and full selector training-data overlap are not independently established. No adaptive attacker, forged verification, or arbitrary downstream-query guarantee is evaluated.

8 Conclusion

We construct a resource attribution framework for RAG recovery by connecting decision-theoretic information comparisons to explicit selection, message, and rendering interfaces. The framework yields matched controls that change the interpretation of observed gains: substantial oracle localization and presentation benefits need not be additional verification-display benefits, nor do they automatically transfer to an answer-blind selector. Controlled claim packets, retrieved-quotation comparisons, and coverage-aware decisions provide complementary evidence under this framework. The contribution is a formal evaluation method and its empirical consequences, with specified information boundaries and outcome objectives.

AI use statement

In this study, generative AI tools were used to generate synthetic data sets, implement relevant methods (statistical-figure generation and data-export scripts), and polish the language and check grammar.

Reproducibility statement

Anonymous supplementary code, frozen results, and figure data are available at `https://anonymous.4open.science/r/fsws62dwas212fdf345`. The archived experimental results can be inspected without installation, training, a GPU, or an API key. The artifact retains frozen inputs, source IDs and quotations, exclusions, model/prompt provenance, raw responses and failures, and scoring/aggregation scripts. Appendices B, C, D, and E specify protocols and offline analyses. Saved responses support reaggregation without new inference; logical views and actual calls are distinguished. Server execution records remain in the author workspace.

REFERENCES

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032. URL https://doi.org/10.1214/aoms/1177729032.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006. doi: 10.1002/047174882X. URL https://doi.org/10.1002/047174882X.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://www.jmlr.org/papers/v11/el-yaniv10a.html.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. LightRAG: Simple and fast retrieval-augmented generation. arXiv preprint arXiv:2410.05779, 2024. URL https://arxiv.org/abs/2410.05779.

Jinpu Jiang, Xuan Wu, Wenhao Song, Bo Yang, You Zhou, Hongwei Ge, Heow Pueh Lee, Yanchun Liang, and Chunguo Wu. ReliableRAG: Combating misinformation in retrieval-augmented generation via reliability-guided reasoning chains. arXiv:2608.25487v1, 2026. URL https://arxiv.org/abs/2608.25487.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.06037. arXiv:2411.06037v3.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.nips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023. URL https://arxiv.org/abs/2307.03172v3. Version 3.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. FaithEval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.03727. arXiv:2410.03727v3.

Mistral AI. Model card for Mistral-7B-Instruct-v0.3. Official model card, 2024. URL https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

Havva Alizadeh Noughabi, Fattane Zarrinkalam, and Ali Dehghantanha. Defense against knowledge poisoning attack on GraphRAG. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 555–563. Association for Computational Linguistics, July 2026. doi: 10.18653/v1/2026.acl-short.47. URL https://aclanthology.org/2026.acl-short.47/.

Qwen Team. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Zeyu Shen, Basileal Imana, Tong Wu, Chong Xiang, Prateek Mittal, and Aleksandra Korolova. ReliabilityRAG: Effective and provably robust defense for RAG-based web-search. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2509.23519. arXiv:2509.23519v2.

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

Unsloth. Qwen3.8-27B-GGUF: Standard Q6_K checkpoint. Hugging Face model repository, 2026. URL https://huggingface.co/unsloth/Qwen3.8-27B-GGUF/tree/f1bfb127c64f7072bdd2cad55f258b9c8b2910fe. Pinned revision f1bfb127c64f7072bdd2cad55f258b9c8b2910fe.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust RAG against retrieval corruption, 2024. URL https://arxiv.org/abs/2405.15556v2. Version 2, revised April 2026.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with compression and selective augmentation, 2023. URL https://arxiv.org/abs/2310.04408v1. Version 1.

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/.

Baolei Zhang, Haoran Xin, Minghong Fang, Zhuqing Liu, Biao Yi, Tong Li, and Zheli Liu. Traceback of poisoning attacks to retrieval-augmented generation. In *The Web Conference*, 2025. URL https://arxiv.org/abs/2504.21668. arXiv:2504.21668v2.

Huichi Zhou, Kin-Hei Lee, Zhonghao Zhan, Yue Chen, Zhenhao Li, Zhaoyang Wang, Hamed Haddadi, and Emine Yilmaz. TrustRAG: Enhancing robustness and trustworthiness in retrieval-augmented generation, 2025. URL https://arxiv.org/abs/2501.00879v3. Version 3.

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

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844. USENIX Association, 2025. URL https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.

A RESOURCE ATTRIBUTION: ASSUMPTIONS, PROOFS, AND SCOPE

A.1 OBSERVATION-AWARE DECOMPOSITION

All expectations use one specified population law with well-defined conditional expectations. We take a finite action set and fixed bounded task loss ℓ; acquisition and computation costs are recorded separately. The construction permits the same visible evidence in different target states. It is not a claim that empirical conditional entropy over 200 uniquely identified questions measures uncertainty about their answers. Shared background and protocol knowledge must be treated consistently across arms; evaluation-label-fitted parameters cannot silently be included in the answer-blind boundary.

For arm j, the disclosed resource is W_j, the actual rendered observation is X_j, and the generator produces an action from X_j. A randomized rendering satisfies $Y \perp X_j \mid W_j$; the generator's randomness is likewise target-independent conditional on X_j. For any decision rule using X_j, a rule with access to W_j can simulate the renderer and then that rule. Randomization cannot improve the minimum conditional expected loss over the finite action set. Thus $R^*(W_j) \le R^*(X_j)$ and $C_j \ge 0$. Bayes optimality on X_j gives $E_j \ge 0$. If W_j is also recoverable from X_j, the reverse simulation gives equality and $C_j = 0$.

If $Y \perp S_b \mid Z$, then for every action a,

$$\mathbb{E}[\ell(Y, a) \mid Z, S_b] = \mathbb{E}[\ell(Y, a) \mid Z].$$

Taking a minimum over actions and an outer expectation gives $R^*(Z, S_b) = R^*(Z)$, and conditional independence gives $I(Y; S_b \mid Z) = 0$. Observers receiving additional S_o or M may ignore them, so $A_S, A_M \ge 0$. Finally,

$$R_g(j) = R^*(W_j) + C_j + E_j$$

implies Proposition 1 by subtraction. In particular,

$$R_g(b) - R_g(u) = A_S + H_b - H_u.$$

The sign of the measured gap cannot identify the magnitude, or strict positivity, of A_S. This is a risk decomposition over specified interfaces, not a procedure for estimating its latent terms from model accuracy.

A.2 A FIXED MARKER VERSUS A VARIABLE MESSAGE

Suppose $X_v = h(X_u)$ and $X_u = k(X_v)$ almost surely for fixed measurable maps. Each observation can simulate every decision rule available to the other, yielding $R^*(X_u) = R^*(X_v)$. Subtracting $R_g(j) = R^*(X_j) + E_j$ proves Proposition 2.

In the D-U/D-V protocol, the oracle-selected quotation is already explicitly identified in the appended slot. The map changes its UNVERIFIED token to VERIFIED, and the inverse restores it. For the nine missing-quotation groups, both maps are the identity. This argument assumes a common known selection/message regime and no truncation. It is a comparison of ideal observers who know that regime; the fixed LLM need not know or exploit it. The marker can affect that LLM's interpretation despite the ideal equivalence.

This result is narrower than saying that verification supplies no information. If a real verifier produces positive, negative, or uncertain messages, those messages need not be recoverable from the original input. Their conditional information and decision value can be positive. Likewise, marking an otherwise unidentified true claim can reveal its identity; stripping all marks need not admit an inverse without privileged information. Claim selection in B/C must therefore not be treated as a fixed-marker rewrite merely because the displayed positive token is always the same. Hidden adaptive selection, lost source binding, truncation, or differing background knowledge can invalidate the stated assumptions and require checking the actual rendered observations.

A.3 COUNTEREXAMPLES AND INTERPRETIVE BOUNDARIES

Oracle location carries information without new vocabulary. Let $Z = z_0$ contain the candidate strings 0 and 1, and let Y be a uniform bit. Let $S_o = Y$ select the corresponding location. Then $I(Y; S_o \mid Z) = 1$ bit and optimal forced binary-answer error falls from $1/2$ to zero, although the

displayed characters already occurred in the context. Information is a property of the specified joint law, not whether a fixed generator can decode the cue. Access to privileged annotations alone does not ensure positive information: the policy may ignore them or the target may already be determined by Z.

Positive information without a 0–1 decision gain. Let Z, S be constant, let $P(M = 0) = P(M = 1) = 1/2$, and set

$$P(Y = 1 \mid M = 0) = 0.6, \qquad P(Y = 1 \mid M = 1) = 0.9.$$

The prior is $P(Y = 1) = 0.75$. The optimal binary classification action is always 1, with error 0.25 before and after observing M, yet $I(Y; M) > 0$. Equal accuracy therefore cannot establish information redundancy. Still less can a small, uncertain EM difference for a nonoptimal language model establish it.

Equal observed gains, different information advantages. Consider forced binary answering with 0–1 loss, a uniform target Y, and lossless rendering. In family I, Z is a constant context listing both candidates, the oracle index is $S_o = Y$, the native rule always answers 0, and the quote-conditioned rule reads the index. Both rules are Bayes optimal for their respective observations: risks are $1/2$ and 0, $A_S = 1/2$, and $H_d - H_u = 0$. In family II, the visible context additionally contains an unambiguous field encoding Y; the same native rule ignores that field, while selecting its value and re-exposing it enables the quote-conditioned rule to answer correctly. Now $R^*(Z) = R^*(Z, S_o) = 0$, so $A_S = 0$ and $H_d - H_u = 1/2$. The observed risk reduction is $1/2$ in both families. The construction illustrates why an accuracy gain alone does not numerically identify its information component; it is an analytic example, not a model experiment or a claim that our dataset follows either family.

Compression. If an answer-blind compressor replaces Z by W, then $R^*(Z) \leq R^*(W)$ by simulation. Equality is not assured when information is discarded. This distinguishes the retained-context quotation interface from a compression-and-replacement system such as RECOMP.

Identified display estimand. For a fixed selector π_o, a prespecified bounded outcome u_i, and matched selected span s_i and message regime m_i, the reported display contrast has the form

$$\widehat{\tau}_T^{\pi_o} = \frac{1}{n} \sum_i \left[u_i(g(\Phi_1(z_i, s_i, m_i))) - u_i(g(\Phi_0(z_i, s_i, m_i))) \right].$$

It conditions on the selector and observed regime. It does not identify performance with erroneous messages or a different selection policy, estimate mutual information, or uniquely allocate effects across intervention orders. Historical baseline reuse also leaves numerical run-to-run variation unmeasured. When u_i is reference matching, the estimand concerns reference recovery, not comprehensive semantic correctness or appropriate abstention.

Relation to the classical foundations. The simulation ordering follows the comparison-of-experiments perspective of Blackwell (Blackwell, 1953); conditional independence and information-processing distinctions follow standard information theory (Cover & Thomas, 2006). Our contribution is the RAG resource interface, its explicit separation of privileged selection from displayed verification, and the resulting matched evaluation construction. The excess-risk identity and data-processing principles are credited to these foundations rather than claimed as new general results.

B CONFIRMATION PROTOCOL AND REPRODUCIBILITY

Selection and provenance. The new sample is constructed before formal answering from existing local copies of the source datasets. Candidate chains require two supporting documents, a literal bridge, and a final-answer-bearing successor passage. MuSiQue uses its two-hop decomposition; HotpotQA starts from a support-linked bridge/cloze relation. Candidates are ordered by a fixed hash and selected greedily with disjoint supporting documents. Previous query IDs and normalized document titles are excluded. Of 500 automatically checked candidates, 164/200 HotpotQA and 141/300 MuSiQue candidates pass the initial semantic checks; subsequent fixed alias, length, and source constraints determine the first 100 eligible groups per dataset. No correctness or attack-success feedback is used in these steps. All rejection and failure records are retained.

Each retained group has two extracted claims and two exact source quotations, giving 400 distinct support documents and 400 quotation spans. Exact-substring checks verify that a quoted span occurs in its source; this does not prove the extracted claim preserves every semantic qualifier. A predetermined 12-group sample received additional automated inspection, not human adjudication. A 20-batch target-type pass and a further 20-batch replacement-subtype pass were completed before test inference. The construction model corrected twenty-seven rare type assignments; these are not human labels; six lexically invalid replacement proposals retain their predefined fallbacks. No cases are removed after observing formal predictions.

Known preprocessing usage is 396751 tokens: 347995 in source annotation, 20534 in target typing, and 28222 in replacement repair. Six initial failed annotation calls lack response usage, making this a lower bound. These are construction costs, separate from test inference. The public benchmark sources and their existing reference answers are not evidence of model pretraining disjointness.

Claim interfaces. Claim identifiers are opaque, with no truth or role labels in the input. The context begins with the following fixed policy:

> Verification applies only to the marked claim, never to its entire source or related claims. VERIFIED claims are externally checked for this benchmark snapshot. UNVERIFIED claims are ordinary evidence and may be incorrect. Resolve a direct conflict using a verified claim when available; otherwise do not assume a source ID or its position indicates truth.

The answer instruction requests the final relation using supplied evidence only, a short answer with no explanation or citations, and `NA` when no definite answer is justified. Each output has exactly one nonempty string field, `answer`. Review receives the identical question and evidence plus the fixed candidate, and may keep, correct, or abstain using the same schema. This avoids the earlier pilot's accept-verbatim contract, which could reject harmless punctuation changes. JSON format correction is allowed at most once; truncated answers are not parsed as successful partial outputs. Transport failures stop the run rather than populating the remaining dataset with failed predictions.

The Sufficient Context prompt is preserved from the authors' repository at commit `801b465eb3be5bf5216fbe3301e61daee85ff99c`, README section "Sufficient Context Autorater Prompt." Timestamp placeholders refer to the source snapshot. The final integer-valued `Sufficient Context` judgment determines retention. It is a prompt adaptation run with the tested local generator; it does not reproduce the original autorater model or train the paper's logistic selection head.

Output-budget extension. Frequent truncation of the explanatory gate output was observed while the first main generator was running. The 768-token results remain immutable. Before additional gate calls, a separate protocol specifies one identical-prompt retry at 4096 tokens for every truncated judgment with a valid original candidate, across both generators and B0/B1. Successful initial judgments and other failures are reused unchanged; no correctness labels select retries. The extension yields 800 logical decision views, not 800 new samples or necessarily 800 new calls. Its input manifest binds the original predictions, calls, prompts, and prepared packets. Tables report the original gate and the completion policy separately, including unresolved failures and the initial-plus-additional cost.

Models and execution. Qwen3-32B-AWQ uses vLLM 0.11.2 and its original pinned checkpoint. A storage I/O failure in the fourth shard was resolved before formal prediction by restoring that shard

into a separate directory, verifying its SHA-256 against the original manifest and the repository's pinned LFS digest. The standard Qwen3.8-27B Q6_K checkpoint is from `unsloth/Qwen3.8-27B-GGUF`, revision `f1bfb127c64f7072bdd2cad55f258b9c8b2910fe`; it is not the locally installed modified model. Its measured SHA-256 matches the repository manifest. It runs through the existing llama.cpp binary, whose hash and runtime information are retained. Both use the same prompt, temperature, seed, per-request context budget, and four-client concurrency. Services run sequentially on one RTX 5090 with approximately 32 GB VRAM and are stopped after their owned runs. No server latency comparison is used as a model-quality claim.

Dense retrieval. Each corpus contains 400 documents and uses normalized MiniLM-L6-v2 embeddings on CPU. Ties are resolved by document ID. Retrieval returns top four document prefixes with their ordinary provenance, plus the same separately verified first-hop claim. The bridge variant appends the verified bridge entity to the original question. Diagnostics distinguish retrieving the correct successor document from exposing its supporting quotation or a reference value in the truncated prefix. A retrieved poison note may belong to a different group. The uniform title style of synthetic poison notes and small isolated corpora limit attack realism. Neither dense variant rebuilds a graph, and no effect is attributed to the graph-versus-flat distinction.

Scoring and paired analysis. The original frozen alias normalization and scoring implementation is reused. A valid answer matching a reference alias is an EM success; strict `NA` and interface failure remain distinct, and failures stay in all primary denominators. Reference-value matches are not claims of complete factuality or evidential faithfulness. Replacement matches may be reported as substring matches; the finer paired transitions separately identify exact replacement answers. Aggregation pairs model, condition, order, method, and group ID. Bootstrap resampling operates within each dataset at the group level; the two datasets each contribute 100 groups. We report exploratory intervals, not population-wide guarantees or multiplicity-adjusted discovery tests. The two retrieval variants produce identical generator prompts for 24 queries; their cached responses account for the difference between 5040 logical main records and 4992 actual responses across both generators.

Limits of exact matching. A small inspection of nonmatching Qwen3.8 C1 outputs found reference/answer pairs such as "Point Place"/"Point Place, Wisconsin," "Tel Aviv"/"Maccabi Tel Aviv," and "in Singapore"/"Singapore." These examples illustrate why a residual nonmatch cannot be called a factual error. No aliases are added and no primary labels are changed after inspecting outputs. Token F1 and its paired resource contrasts provide a frozen secondary view; neither metric substitutes for comprehensive semantic validation.

Cross-family addition. The final core-only replication uses the earlier pinned Mistral-7B-Instruct-v0.3 checkpoint, revision `c170c708c41dac9275d15a8fff4eca08d52bab71`, with BF16 weights and vLLM 0.11.2. File hashes are checked against its earlier download manifest. The exact main inference implementation is reused for the 1200 direct calls. This addition was specified after Qwen3.8 core results and before Qwen3-32B completion; no Mistral output was used to revise inputs or prompts. It is not retrospectively labeled as part of the original prespecified two-generator experiment.

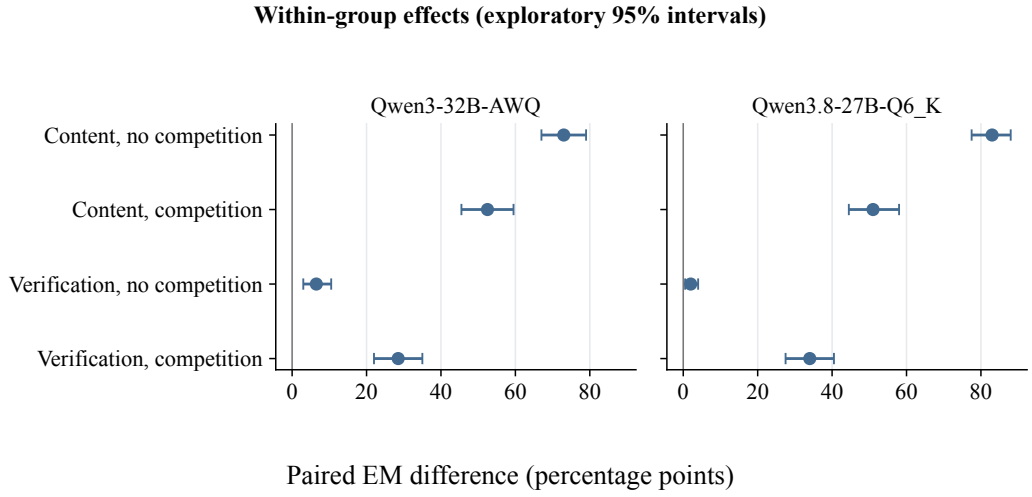

Figure 4: Paired content and verification effects for the original two-generator confirmation. Intervals use 2000 group resamples stratified by dataset and are exploratory. The verification comparison holds claim text fixed and adds an oracle verification flag.

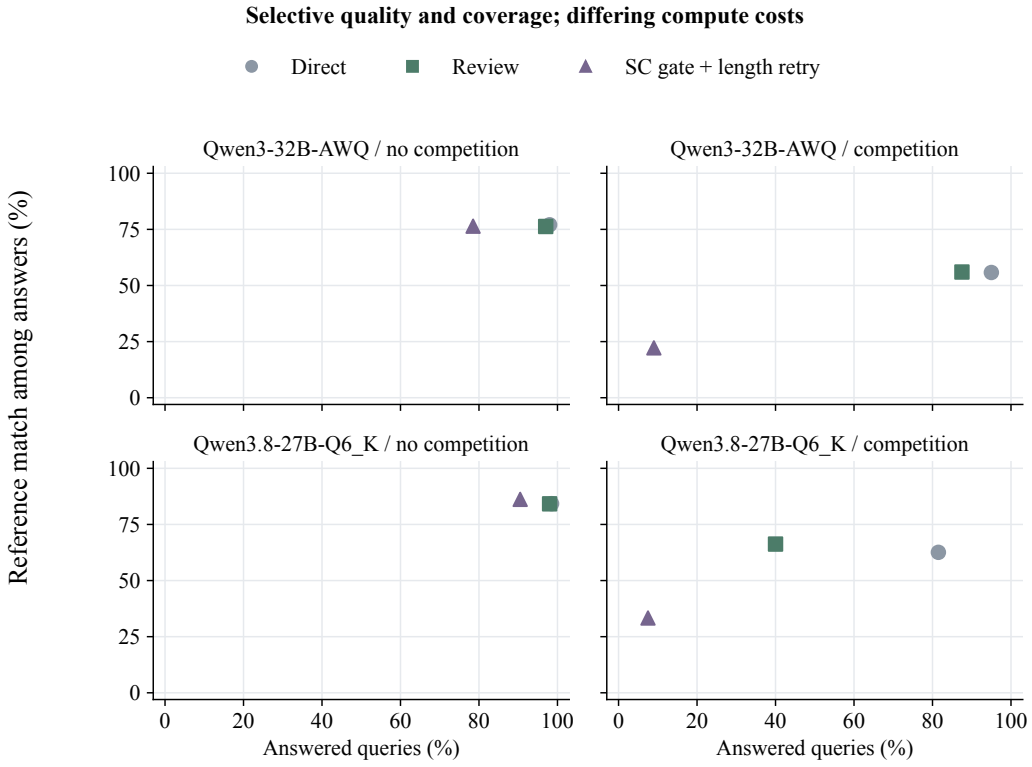

Figure 5: Quality and coverage at each observed decision operating point. Quality is frozen reference matching among valid non-NA answers, and is undefined at zero coverage. Points use different inference costs; they are not a threshold-swept frontier.

Table 4: Complete reference-based outcomes: all 200 groups.

Model	Condition	Method	EM count	F1%	NA	Fail	Repl.	Ans.	Quality%
Qwen3-32B	A0	Direct	5	8.2	113	0	1	87	5.7
Qwen3-32B	A1	Direct	1	14.7	12	0	147	188	0.5
Qwen3-32B	B0	Direct	151	86.4	4	0	1	196	77.0
Qwen3-32B	B0	Review	148	84.9	6	0	1	194	76.3
Qwen3-32B	B0	SC-768	107	61.6	35	25	0	140	76.4
Qwen3-32B	B1	Direct	106	64.4	10	0	45	190	55.8
Qwen3-32B	B1	Review	98	60.4	25	0	38	175	56.0
Qwen3-32B	B1	SC-768	2	2.6	149	40	0	11	18.2
Qwen3-32B	C0	Direct	164	91.3	0	0	1	200	82.0
Qwen3-32B	C1	Direct	163	90.4	2	0	3	198	82.3
Qwen3-32B	dense_bridge	Direct	104	64.5	3	0	55	197	52.8
Qwen3-32B	dense_query	Direct	70	47.8	21	0	53	179	39.1
Qwen3-32B	B0	SC-completed	120	68.8	43	0	0	157	76.4
Qwen3-32B	B1	SC-completed	4	4.5	182	0	1	18	22.2
Qwen3.8-27B	A0	Direct	0	1.3	188	0	0	12	0.0
Qwen3.8-27B	A1	Direct	0	16.0	8	0	176	192	0.0
Qwen3.8-27B	B0	Direct	166	92.7	3	0	1	197	84.3
Qwen3.8-27B	B0	Review	165	92.2	4	0	1	196	84.2
Qwen3.8-27B	B0	SC-768	107	58.1	6	71	0	123	87.0
Qwen3.8-27B	B1	Direct	102	60.8	37	0	38	163	62.6
Qwen3.8-27B	B1	Review	53	31.5	120	0	11	80	66.3
Qwen3.8-27B	B1	SC-768	0	0.0	88	111	0	1	0.0
Qwen3.8-27B	C0	Direct	170	93.9	2	0	1	198	85.9
Qwen3.8-27B	C1	Direct	170	93.5	1	0	2	199	85.4
Qwen3.8-27B	dense_bridge	Direct	102	66.0	7	0	55	193	52.8
Qwen3.8-27B	dense_query	Direct	73	48.2	37	0	53	163	44.8
Qwen3.8-27B	B0	SC-completed	156	85.8	19	0	0	181	86.2
Qwen3.8-27B	B1	SC-completed	5	5.0	185	0	2	15	33.3

EM is a count; F1 and answered-reference quality are percentages. Repl. is replacement-value matching, not an independent semantic error judgment. Failures stay in the denominator. SC-completed retains all valid initial judgments and retries only truncations.

Table 5: Complete reference-based outcomes: 100 HotpotQA groups.

Model	Condition	Method	EM count	F1%	NA	Fail	Repl.	Ans.	Quality%
Qwen3-32B	A0	Direct	3	8.5	57	0	1	43	7.0
Qwen3-32B	A1	Direct	0	8.3	6	0	77	94	0.0
Qwen3-32B	B0	Direct	71	84.3	2	0	1	98	72.4
Qwen3-32B	B0	Review	70	83.3	2	0	1	98	71.4
Qwen3-32B	B0	SC-768	54	63.6	14	14	0	72	75.0
Qwen3-32B	B1	Direct	54	64.8	3	0	21	97	55.7
Qwen3-32B	B1	Review	51	62.3	9	0	18	91	56.0
Qwen3-32B	B1	SC-768	1	2.6	71	25	0	4	25.0
Qwen3-32B	C0	Direct	75	87.1	0	0	1	100	75.0
Qwen3-32B	C1	Direct	76	87.6	1	0	2	99	76.8
Qwen3-32B	dense_bridge	Direct	55	68.5	1	0	21	99	55.6
Qwen3-32B	dense_query	Direct	43	54.9	6	0	20	94	45.7
Qwen3-32B	B0	SC-completed	60	70.9	18	0	0	82	73.2
Qwen3-32B	B1	SC-completed	1	4.5	92	0	1	8	12.5
Qwen3.8-27B	A0	Direct	0	2.3	91	0	0	9	0.0
Qwen3.8-27B	A1	Direct	0	10.5	3	0	86	97	0.0
Qwen3.8-27B	B0	Direct	77	89.5	2	0	1	98	78.6
Qwen3.8-27B	B0	Review	76	88.5	3	0	1	97	78.4
Qwen3.8-27B	B0	SC-768	55	59.6	4	32	0	64	85.9
Qwen3.8-27B	B1	Direct	53	63.3	17	0	13	83	63.9
Qwen3.8-27B	B1	Review	23	28.8	60	0	6	40	57.5
Qwen3.8-27B	B1	SC-768	0	0.0	37	63	0	0	–
Qwen3.8-27B	C0	Direct	78	90.1	2	0	1	98	79.6
Qwen3.8-27B	C1	Direct	79	90.4	1	0	1	99	79.8
Qwen3.8-27B	dense_bridge	Direct	57	72.2	2	0	18	98	58.2
Qwen3.8-27B	dense_query	Direct	42	54.1	14	0	19	86	48.8
Qwen3.8-27B	B0	SC-completed	71	80.6	13	0	0	87	81.6
Qwen3.8-27B	B1	SC-completed	2	5.5	92	0	0	8	25.0

EM is a count; F1 and answered-reference quality are percentages. Repl. is replacement-value matching, not an independent semantic error judgment. Failures stay in the denominator. SC-completed retains all valid initial judgments and retries only truncations.

Table 6: Complete reference-based outcomes: 100 MuSiQue groups.

Model	Condition	Method	EM count	F1%	NA	Fail	Repl.	Ans.	Quality%
Qwen3-32B	A0	Direct	2	7.9	56	0	0	44	4.5
Qwen3-32B	A1	Direct	1	21.1	6	0	70	94	1.1
Qwen3-32B	B0	Direct	80	88.6	2	0	0	98	81.6
Qwen3-32B	B0	Review	78	86.6	4	0	0	96	81.3
Qwen3-32B	B0	SC-768	53	59.6	21	11	0	68	77.9
Qwen3-32B	B1	Direct	52	64.0	7	0	24	93	55.9
Qwen3-32B	B1	Review	47	58.4	16	0	20	84	56.0
Qwen3-32B	B1	SC-768	1	2.6	78	15	0	7	14.3
Qwen3-32B	C0	Direct	89	95.6	0	0	0	100	89.0
Qwen3-32B	C1	Direct	87	93.3	1	0	1	99	87.9
Qwen3-32B	dense_bridge	Direct	49	60.6	2	0	34	98	50.0
Qwen3-32B	dense_query	Direct	27	40.7	15	0	33	85	31.8
Qwen3-32B	B0	SC-completed	60	66.6	25	0	0	75	80.0
Qwen3-32B	B1	SC-completed	3	4.6	90	0	0	10	30.0
Qwen3.8-27B	A0	Direct	0	0.3	97	0	0	3	0.0
Qwen3.8-27B	A1	Direct	0	21.5	5	0	90	95	0.0
Qwen3.8-27B	B0	Direct	89	95.8	1	0	0	99	89.9
Qwen3.8-27B	B0	Review	89	95.8	1	0	0	99	89.9
Qwen3.8-27B	B0	SC-768	52	56.6	2	39	0	59	88.1
Qwen3.8-27B	B1	Direct	49	58.4	20	0	25	80	61.3
Qwen3.8-27B	B1	Review	30	34.3	60	0	5	40	75.0
Qwen3.8-27B	B1	SC-768	0	0.0	51	48	0	1	0.0
Qwen3.8-27B	C0	Direct	92	97.6	0	0	0	100	92.0
Qwen3.8-27B	C1	Direct	91	96.6	0	0	1	100	91.0
Qwen3.8-27B	dense_bridge	Direct	45	59.9	5	0	37	95	47.4
Qwen3.8-27B	dense_query	Direct	31	42.3	23	0	34	77	40.3
Qwen3.8-27B	B0	SC-completed	85	91.1	6	0	0	94	90.4
Qwen3.8-27B	B1	SC-completed	3	4.4	93	0	2	7	42.9

EM is a count; F1 and answered-reference quality are percentages. Repl. is replacement-value matching, not an independent semantic error judgment. Failures stay in the denominator. SC-completed retains all valid initial judgments and retries only truncations.

Table 7: Paired resource and retrieval contrasts: both datasets.

Model	Change	EM pp [95% CI]	F1 pp [95% CI]
Qwen3-32B	A0→B0	+73.0 [+67.0, +79.0]	+78.2 [+73.3, +83.0]
Qwen3-32B	A1→B1	+52.5 [+45.5, +59.5]	+49.7 [+43.6, +55.9]
Qwen3-32B	B0→C0	+6.5 [+3.0, +10.5]	+4.9 [+2.0, +8.0]
Qwen3-32B	B1→C1	+28.5 [+22.0, +35.0]	+26.0 [+20.8, +31.6]
Qwen3-32B	A0→A1	-2.0 [-4.0, -0.5]	+6.5 [+3.1, +9.9]
Qwen3-32B	B0→B1	-22.5 [-29.0, -16.5]	-22.0 [-27.4, -17.0]
Qwen3-32B	C0→C1	-0.5 [-4.5, +3.5]	-0.9 [-3.5, +1.7]
Qwen3-32B	dense_query→dense_bridge	+17.0 [+10.5, +23.5]	+16.7 [+11.1, +22.7]
Qwen3.8-27B	A0→B0	+83.0 [+77.5, +88.0]	+91.3 [+88.1, +94.3]
Qwen3.8-27B	A1→B1	+51.0 [+44.5, +58.0]	+44.9 [+38.1, +51.5]
Qwen3.8-27B	B0→C0	+2.0 [+0.5, +4.0]	+1.2 [+0.2, +2.6]
Qwen3.8-27B	B1→C1	+34.0 [+27.5, +40.5]	+32.7 [+26.5, +38.9]
Qwen3.8-27B	A0→A1	+0.0 [+0.0, +0.0]	+14.7 [+11.6, +17.9]
Qwen3.8-27B	B0→B1	-32.0 [-39.0, -25.0]	-31.8 [-37.8, -25.5]
Qwen3.8-27B	C0→C1	+0.0 [-3.0, +3.0]	-0.4 [-2.1, +1.1]
Qwen3.8-27B	dense_query→dense_bridge	+14.5 [+8.0, +20.5]	+17.8 [+11.8, +23.8]

Exploratory group bootstrap, 2000 resamples within dataset. Differences are in percentage points.

Table 8: Paired resource and retrieval contrasts: hotpotqa.

Model	Change	EM pp [95% CI]	F1 pp [95% CI]
Qwen3-32B	A0→B0	+68.0 [+58.0, +77.0]	+75.8 [+68.5, +83.1]
Qwen3-32B	A1→B1	+54.0 [+44.0, +64.0]	+56.5 [+47.8, +65.0]
Qwen3-32B	B0→C0	+4.0 [+0.0, +9.0]	+2.8 [-1.5, +7.2]
Qwen3-32B	B1→C1	+22.0 [+14.0, +31.0]	+22.8 [+15.8, +30.3]
Qwen3-32B	A0→A1	-3.0 [-7.0, +0.0]	-0.2 [-5.0, +4.7]
Qwen3-32B	B0→B1	-17.0 [-25.0, -9.0]	-19.5 [-27.2, -12.4]
Qwen3-32B	C0→C1	+1.0 [-4.0, +6.0]	+0.5 [-4.1, +5.0]
Qwen3-32B	dense_query→dense_bridge	+12.0 [+3.0, +21.0]	+13.6 [+5.0, +22.0]
Qwen3.8-27B	A0→B0	+77.0 [+68.0, +85.0]	+87.2 [+81.5, +92.1]
Qwen3.8-27B	A1→B1	+53.0 [+43.0, +63.0]	+52.9 [+43.5, +61.9]
Qwen3.8-27B	B0→C0	+1.0 [+0.0, +3.0]	+0.7 [+0.0, +1.8]
Qwen3.8-27B	B1→C1	+26.0 [+18.0, +35.0]	+27.1 [+18.8, +35.5]
Qwen3.8-27B	A0→A1	+0.0 [+0.0, +0.0]	+8.1 [+4.7, +11.6]
Qwen3.8-27B	B0→B1	-24.0 [-33.0, -15.0]	-26.2 [-35.0, -18.0]
Qwen3.8-27B	C0→C1	+1.0 [-3.0, +5.0]	+0.3 [-2.5, +2.8]
Qwen3.8-27B	dense_query→dense_bridge	+15.0 [+8.0, +23.0]	+18.0 [+9.9, +26.7]

Exploratory group bootstrap, 2000 resamples within dataset. Differences are in percentage points.

Table 9: Paired resource and retrieval contrasts: musique.

Model	Change	EM pp [95% CI]	F1 pp [95% CI]
Qwen3-32B	A0→B0	+78.0 [+70.0, +86.0]	+80.7 [+73.8, +86.8]
Qwen3-32B	A1→B1	+51.0 [+42.0, +61.0]	+43.0 [+34.0, +52.3]
Qwen3-32B	B0→C0	+9.0 [+4.0, +15.0]	+7.0 [+2.7, +11.8]
Qwen3-32B	B1→C1	+35.0 [+26.0, +45.0]	+29.2 [+21.2, +37.3]
Qwen3-32B	A0→A1	-1.0 [-3.0, +0.0]	+13.1 [+8.0, +18.0]
Qwen3-32B	B0→B1	-28.0 [-38.0, -18.0]	-24.6 [-32.2, -17.1]
Qwen3-32B	C0→C1	-2.0 [-7.0, +3.0]	-2.3 [-5.7, +0.1]
Qwen3-32B	dense_query→dense_bridge	+22.0 [+13.0, +32.0]	+19.9 [+11.9, +27.7]
Qwen3.8-27B	A0→B0	+89.0 [+82.0, +95.0]	+95.5 [+92.1, +98.2]
Qwen3.8-27B	A1→B1	+49.0 [+39.0, +59.0]	+36.9 [+26.9, +47.0]
Qwen3.8-27B	B0→C0	+3.0 [+0.0, +7.0]	+1.8 [+0.0, +4.2]
Qwen3.8-27B	B1→C1	+42.0 [+32.0, +52.0]	+38.2 [+29.5, +47.5]
Qwen3.8-27B	A0→A1	+0.0 [+0.0, +0.0]	+21.2 [+16.2, +26.4]
Qwen3.8-27B	B0→B1	-40.0 [-51.0, -30.0]	-37.4 [-47.1, -28.5]
Qwen3.8-27B	C0→C1	-1.0 [-5.0, +2.0]	-1.0 [-3.0, +0.4]
Qwen3.8-27B	dense_query→dense_bridge	+14.0 [+5.0, +23.0]	+17.7 [+9.2, +26.3]

Exploratory group bootstrap, 2000 resamples within dataset. Differences are in percentage points.

Table 10: Predetermined order replay on 40 groups.

Model	Condition	Main EM	Swapped EM	Gains	Losses	Change pp [95% CI]
Qwen3-32B	A1	0	0	0	0	+0.0 [+0.0, +0.0]
Qwen3-32B	B1	23	19	6	10	-10.0 [-30.0, +7.6]
Qwen3-32B	C1	31	33	3	1	+5.0 [-5.0, +15.0]
Qwen3.8-27B	A1	0	0	0	0	+0.0 [+0.0, +0.0]
Qwen3.8-27B	B1	26	20	5	11	-15.0 [-35.0, +5.0]
Qwen3.8-27B	C1	33	32	0	1	-2.5 [-7.5, +0.0]

Twenty groups per dataset; these replays do not add independent observations.

Table 11: Document and visible-prefix retrieval diagnostics.

Dataset	Query	n	Successor doc	Exact quote	Reference	Any poison
Both	query	200	121	120	119	159
Both	bridge	200	192	191	190	194
hotpotqa	query	100	68	68	67	69
hotpotqa	bridge	100	98	98	97	97
musique	query	100	53	52	52	90
musique	bridge	100	94	93	93	97

Exact quote and reference values are checked within the retrieved successor document prefix. Any poison may be from another group.

Table 12: Completed gate versus random retention at its observed coverage.

Model	Condition	Retained	Gate EM	Random expected EM	Difference
Qwen3-32B	B0	157	120	120.95	-0.95
Qwen3-32B	B1	18	4	10.04	-6.04
Qwen3.8-27B	B0	181	156	152.52	+3.48
Qwen3.8-27B	B1	15	5	9.39	-4.39

The comparison randomly retains the same number of valid non-NA direct candidates. It is an analytic expectation, not another trained defense. Zero coverage has undefined selective quality.

Table 13: Actual logged inference responses and usage.

Model	Stage	Responses	Total tokens	Truncations
Qwen3-32B	Main	2496	1641263	65
Qwen3-32B	Gate completion	65	117817	0
Qwen3.8-27B	Main	2496	1681605	182
Qwen3.8-27B	Gate completion	182	378064	0
Mistral-7B	Core replication	1200	576203	0

Construction costs are separate. Shared cache reuse makes actual responses differ from logical records.

Table 14: Mistral core-only cross-family replication.

Dataset	Condition	EM count	F1%	NA	Fail	Repl.
Both	A0	5	7.0	93	0	1
Both	A1	3	9.2	59	0	73
Both	B0	71	45.3	44	0	0
Both	B1	70	46.8	43	0	24
Both	C0	97	59.0	26	0	0
Both	C1	110	68.0	26	0	2
hotpotqa	A0	5	9.9	39	0	1
hotpotqa	A1	3	8.7	22	0	41
hotpotqa	B0	42	52.0	15	0	0
hotpotqa	B1	37	50.0	18	0	11
hotpotqa	C0	49	60.9	10	0	0
hotpotqa	C1	57	71.4	10	0	0
musique	A0	0	4.1	54	0	0
musique	A1	0	9.7	37	0	32
musique	B0	29	38.5	29	0	0
musique	B1	33	43.5	25	0	13
musique	C0	48	57.0	16	0	0
musique	C1	53	64.6	16	0	2

All 200 existing groups; 100 per dataset. This addition follows the first Qwen core results and adds no new independent groups.

Table 15: Mistral paired resource effects.

Dataset	Change	EM pp [95% CI]	F1 pp [95% CI]
Both	A0→B0	+33.0 [+26.5, +39.5]	+38.2 [+32.0, +44.5]
Both	A1→B1	+33.5 [+27.0, +40.0]	+37.6 [+31.7, +43.6]
Both	B0→C0	+13.0 [+8.0, +18.5]	+13.7 [+8.9, +18.6]
Both	B1→C1	+20.0 [+14.0, +26.0]	+21.3 [+16.1, +27.0]
Both	B0→B1	-0.5 [-6.5, +5.5]	+1.5 [-4.3, +7.2]
Both	C0→C1	+6.5 [+0.0, +13.0]	+9.1 [+3.1, +14.6]
hotpotqa	A0→B0	+37.0 [+28.0, +47.0]	+42.1 [+33.2, +51.0]
hotpotqa	A1→B1	+34.0 [+25.0, +44.0]	+41.3 [+32.1, +49.8]
hotpotqa	B0→C0	+7.0 [+1.0, +13.0]	+8.9 [+2.4, +15.9]
hotpotqa	B1→C1	+20.0 [+12.0, +28.0]	+21.4 [+13.4, +29.5]
hotpotqa	B0→B1	-5.0 [-14.0, +4.0]	-2.0 [-11.0, +7.2]
hotpotqa	C0→C1	+8.0 [-1.0, +17.0]	+10.5 [+2.7, +18.3]
musique	A0→B0	+29.0 [+21.0, +39.0]	+34.3 [+26.4, +43.0]
musique	A1→B1	+33.0 [+24.0, +42.0]	+33.8 [+25.6, +42.5]
musique	B0→C0	+19.0 [+11.0, +28.0]	+18.5 [+11.1, +26.0]
musique	B1→C1	+20.0 [+12.0, +28.0]	+21.1 [+13.7, +28.4]
musique	B0→B1	+4.0 [-4.0, +11.0]	+5.0 [-2.5, +12.6]
musique	C0→C1	+5.0 [-4.0, +14.0]	+7.6 [-0.9, +16.2]

Table 16: Mistral verification-by-competition interaction.

Dataset	EM pp [95% CI]
Both	+7.0 [+0.0, +13.5]
hotpotqa	+13.0 [+4.0, +22.0]
musique	+1.0 [-9.0, +10.0]

The interaction is (C1 minus B1) minus (C0 minus B0).

C RETRIEVED-QUOTATION ATTRIBUTION AND RECOVERY PATHS

This follow-up was specified after the original study and its mock review. It adds no independent fact groups and uses one previously evaluated generator, Qwen3-32B-AWQ. The original query, four retrieved prefixes, their order, first-hop verification and answering instruction are unchanged. Visibility is determined before new answering by exact matching of the frozen supporting quotation in the retrieved document belonging to the group's successor source. Oracle source ownership is used for this localization; it is not an inferred deployed selection policy.

For a visible quotation, both D-U and D-V append the same source-bound text as an additional claim. Its status is UNVERIFIED in D-U and VERIFIED in D-V. All original document claims remain unverified, and only the quoted span receives the new status. A span may contain several facts, so verification cost is not assumed equal to the earlier atomic-claim intervention. For the nine groups without the full quotation in the visible prefix, both inputs equal D and its answer is reused. The primary denominator remains 200, with the precomputed visibility strata reported separately.

The existing inference implementation, checkpoint, serving engine, seed, temperature, answer schema and 128-token budget are reused. There are 400 added condition views: 382 new responses and 18 reused native answers. The 200 native D records are also reused, and no new retrieval is run. Added inference consumes 240258 tokens, with no truncations or interface failures. Native answers come from the previous identical-input configuration; run-to-run numerical nondeterminism is not separately measured. Input and parent hashes, quote positions, raw calls and scoring are retained. Intervals use the same 2000 within-dataset group resamples as the original study.

The B0/B1/C1 path analysis separately reuses all three generators' original atomic-claim outputs. It is a post-hoc descriptive decomposition, not new independent confirmation or an appropriate-abstention labeling exercise.

Table 17: Complete retrieved-quotation outcomes.

Dataset	Subset	Input	n	EM count	F1%	NA	Fail	Repl.
Both	all	D	200	104	64.5	3	0	55
Both	all	D-U	200	135	79.6	2	0	24
Both	all	D-V	200	137	81.8	2	0	16
Both	visible	D	191	104	66.9	1	0	50
Both	visible	D-U	191	135	82.7	0	0	19
Both	visible	D-V	191	137	85.0	0	0	11
Both	missing	D	9	0	13.0	2	0	5
Both	missing	D-U	9	0	13.0	2	0	5
Both	missing	D-V	9	0	13.0	2	0	5
hotpotqa	all	D	100	55	68.5	1	0	21
hotpotqa	all	D-U	100	70	83.3	0	0	7
hotpotqa	all	D-V	100	68	83.5	0	0	5
hotpotqa	visible	D	98	55	69.9	1	0	19
hotpotqa	visible	D-U	98	70	85.0	0	0	5
hotpotqa	visible	D-V	98	68	85.2	0	0	3
hotpotqa	missing	D	2	0	0.0	0	0	2
hotpotqa	missing	D-U	2	0	0.0	0	0	2
hotpotqa	missing	D-V	2	0	0.0	0	0	2
musique	all	D	100	49	60.6	2	0	34
musique	all	D-U	100	65	75.9	2	0	17
musique	all	D-V	100	69	80.1	2	0	11
musique	visible	D	93	49	63.9	0	0	31
musique	visible	D-U	93	65	80.4	0	0	14
musique	visible	D-V	93	69	84.9	0	0	8
musique	missing	D	7	0	16.7	2	0	3
musique	missing	D-U	7	0	16.7	2	0	3
musique	missing	D-V	7	0	16.7	2	0	3

Table 18: Retrieved-quotation paired effects; exploratory stratified group intervals.

Dataset	Subset	Change	EM pp [95% CI]	F1 pp [95% CI]
Both	all	D→D-U	+15.5 [+10.0, +20.5]	+15.1 [+10.8, +19.9]
Both	all	D-U→D-V	+1.0 [-2.0, +4.0]	+2.2 [+0.1, +4.6]
Both	all	D→D-V	+16.5 [+10.5, +22.5]	+17.3 [+12.4, +22.3]
Both	visible	D→D-U	+16.2 [+10.5, +22.0]	+15.8 [+11.2, +20.5]
Both	visible	D-U→D-V	+1.0 [-2.1, +4.2]	+2.3 [+0.1, +4.8]
Both	visible	D→D-V	+17.3 [+11.0, +23.6]	+18.1 [+13.0, +23.4]
Both	missing	D→D-U	+0.0 [+0.0, +0.0]	+0.0 [+0.0, +0.0]
Both	missing	D-U→D-V	+0.0 [+0.0, +0.0]	+0.0 [+0.0, +0.0]
Both	missing	D→D-V	+0.0 [+0.0, +0.0]	+0.0 [+0.0, +0.0]
hotpotqa	all	D→D-U	+15.0 [+8.0, +22.0]	+14.8 [+8.3, +21.7]
hotpotqa	all	D-U→D-V	-2.0 [-6.0, +2.0]	+0.2 [-1.7, +2.7]
hotpotqa	all	D→D-V	+13.0 [+6.0, +21.0]	+15.0 [+8.4, +22.3]
hotpotqa	visible	D→D-U	+15.3 [+8.2, +22.4]	+15.1 [+8.7, +22.4]
hotpotqa	visible	D-U→D-V	-2.0 [-6.1, +2.0]	+0.2 [-1.8, +2.9]
hotpotqa	visible	D→D-V	+13.3 [+6.1, +21.4]	+15.3 [+8.9, +22.3]
hotpotqa	missing	D→D-U	+0.0 [+0.0, +0.0]	+0.0 [+0.0, +0.0]
hotpotqa	missing	D-U→D-V	+0.0 [+0.0, +0.0]	+0.0 [+0.0, +0.0]
hotpotqa	missing	D→D-V	+0.0 [+0.0, +0.0]	+0.0 [+0.0, +0.0]
musique	all	D→D-U	+16.0 [+7.0, +25.0]	+15.3 [+9.1, +21.6]
musique	all	D-U→D-V	+4.0 [-1.0, +9.0]	+4.2 [+0.8, +8.6]
musique	all	D→D-V	+20.0 [+12.0, +29.0]	+19.5 [+12.7, +26.8]
musique	visible	D→D-U	+17.2 [+7.5, +26.9]	+16.5 [+10.0, +23.4]
musique	visible	D-U→D-V	+4.3 [+0.0, +9.7]	+4.5 [+0.9, +8.9]
musique	visible	D→D-V	+21.5 [+12.9, +30.1]	+21.0 [+13.9, +28.5]
musique	missing	D→D-U	+0.0 [+0.0, +0.0]	+0.0 [+0.0, +0.0]
musique	missing	D-U→D-V	+0.0 [+0.0, +0.0]	+0.0 [+0.0, +0.0]
musique	missing	D→D-V	+0.0 [+0.0, +0.0]	+0.0 [+0.0, +0.0]

Table 19: Post-hoc reference-match paths in the original atomic-claim study.

Model	Dataset	n	000	001	010	011	100	101	110	111
qwen32	Both	200	30	16	1	2	6	42	0	103
qwen32	hotpotqa	100	22	6	0	1	2	16	0	53
qwen32	musique	100	8	10	1	1	4	26	0	50
qwen38	Both	200	25	5	2	2	3	65	0	98
qwen38	hotpotqa	100	18	3	1	1	2	24	0	51
qwen38	musique	100	7	2	1	1	1	41	0	47
mistral	Both	200	81	29	0	19	7	13	2	49
mistral	hotpotqa	100	37	12	0	9	5	9	1	27
mistral	musique	100	44	17	0	10	2	4	1	22

Bits indicate EM in B0, B1, and C1. For example, 101 is an original match lost under competition and restored after successor verification; 001 is a match gained from an original nonmatch. Zero does not establish factual error. These are descriptive strata of existing groups.

D FROZEN ANSWER-BLIND SELECTION FOLLOW-UP

The framework consolidates the preceding core and quotation studies; those experiments are not retrospectively claimed as preregistered tests of the formal propositions. This answer-blind follow-up was specified after that analysis and before any D-B responses. It reuses the same 200 groups and generator. We freeze the RECOMP NQ checkpoint and use the authors' mean-pooling and unnormalized dot-product scoring functions (Xu et al., 2023):

Model: `fangyuan/nq_extractive_compressor`
Revision: `46f98c1f57fa919c72d86ced16e3ece3403f9e9e`
Code commit: `51d4432151efb3275257a9407dc71d1e5ec6634d`

The checkpoint and scoring functions are unchanged. Input construction and retained-context application are fixed adapters, not a reproduction of the complete RECOMP system or a comparison against its published scores.

The selector reads only the question, visible claims, and existing verification flags. Its query concatenates the question and visible verified first-hop text. Candidates come from all four prefixes, including polluted and distractor text. The first line is a title; body text is split after sentence-ending punctuation followed by whitespace and an uppercase letter, quote, or opening parenthesis. Each candidate is a literal contiguous prefix of at most 128 generator tokens. Ranking text is the document title plus candidate; highest score wins, with ties resolved in original document/segment order. This deterministic segmentation is not assumed to be a perfect sentence parser. No reference answers, aliases, supporting quotations, source-role labels, or answer outcomes enter selection.

D-B preserves all original claims, order, and statuses and appends one source-bound UNVERIFIED span with D-U's display format. All 200 groups receive the intervention, including nine whose oracle quotation was absent. Oracle quotations have at most 76 tokens and selected blind spans at most 106. The common 128-token cap leaves old oracle inputs unchanged, but does not match each pair's realized length, granularity, or source. The NQ checkpoint is used on both datasets without training, threshold tuning, or outcome-based version selection. This avoids selecting a target-dataset-specific checkpoint but does not establish absence of pretraining or training-text overlap. Query construction, NQ transfer, and single-span selection can all affect the result. Source ownership is used only after selection and inference for diagnosis.

Decision rule and scope. The primary D-B minus D-U contrast uses a practical noninferiority margin of -5 EM points, approximately one third of the earlier 15.5-point oracle gain. It is a study-specific tolerance, not a domain standard. Support for noninferiority requires the lower endpoint of the two-sided, dataset-stratified paired 95% bootstrap interval to exceed -5. The observed interval $[-24, -13]$ lies entirely below it. D-B minus D and secondary outcomes are exploratory. No new selectors, prompts, or samples are added in response to the result.

Table 20: D-B and historical controls. Each dataset contains 100 groups. EM and replacement matches are counts; F1 is a percentage.

Dataset	Input	EM	F1	NA	Replacement
HotpotQA	D	55	68.47	1	21
HotpotQA	D-B	51	64.87	1	21
HotpotQA	D-U	70	83.29	0	7
HotpotQA	D-V	68	83.49	0	5
MuSiQue	D	49	60.56	2	34
MuSiQue	D-B	47	57.64	2	35
MuSiQue	D-U	65	75.89	2	17
MuSiQue	D-V	69	80.09	2	11

For the 191 oracle-visible groups, the EM difference from D-U is -19.37 points $[-25.65, -13.61]$. In the other nine, D-B and D have identical EM/F1 despite new D-B generation; these groups do not explain the gap. Overall F1 differences are -18.34 points $[-23.53, -13.38]$ versus D-U and -3.26 $[-6.57, 0.07]$ versus D. Source diagnostics count 159 selections from the target first-hop document,

Table 21: Paired D-B effects: 2000 within-dataset bootstrap resamples. Gains and losses count reference matches.

Dataset	Comparator	EM diff. (pp)	95% interval	Gains	Losses
All	D	-3.0	[-7.0, 1.0]	5	11
All	D-U	-18.5	[-24.0, -13.0]	2	39
HotpotQA	D	-4.0	[-10.0, 2.0]	3	7
HotpotQA	D-U	-19.0	[-27.0, -12.0]	0	19
MuSiQue	D	-2.0	[-7.0, 3.0]	2	4
MuSiQue	D-U	-18.0	[-27.0, -10.0]	2	20

28 from its successor, nine from its poison note, and four from other groups. These document origins neither establish sentence truth nor identify the causal reason for the gap.

Cost and execution. The selector scores 2199 candidates in approximately 1.56 seconds of GPU scoring time; its measured interval including model loading is approximately 2.52 seconds, excluding checkpoint download and generator startup. Peak allocated GPU memory is approximately 507 MB. Encoder inputs have at most 137 tokens, with no encoder truncation. The weight file is 438,000,173 bytes and its fixed-version hash is retained. These are single-run component costs, not deployment latency. The 200 new generator responses use 125,175 input and 2,117 output tokens, with zero interface failures or truncations. D/D-U/D-V reuse 600 historical views. No new independent groups, retriever runs, or human labels are added; historical reuse leaves numerical run-to-run variation unmeasured.

E POST-HOC DIAGNOSTICS FROM FROZEN RESPONSES

These analyses reuse the saved 200-group predictions and frozen alias scorer; no new model responses or human labels are added. We assign mutually exclusive states in the order interface failure, reference match, NA, designated replacement match, and other nonmatch. The original flags have no overlaps in the 800 D/D-B/D-U/D-V records. A transition into other nonmatch is not evidence of safety or semantic correctness.

For selection diagnostics, document origin is determined only after selection from the saved ownership map. Gains and losses compare D-B with D or D-U on every group; source categories are descriptive and are not randomized interventions. Lexical overlap uses unique tokens after the frozen normalization: first-hop containment is the intersection divided by the number of first-hop tokens; Jaccard divides by the union. Selected spans exclude document titles. No overlap threshold or outcome-dependent subgroup is selected, and these quantities do not establish semantic redundancy or explain a causal mechanism.

Table 22: All D-U → D-V exclusive outcome transitions. Priority is error, reference match, NA, designated replacement match, then other nonmatch; all raw-flag overlaps are zero. Reference uses saved aliases; replacement uses normalized phrase containment.

D-U / D-V	Error	Reference	NA	Replacement	Other	Total
Error	0	0	0	0	0	0
Reference	0	131	0	0	4	135
NA	0	0	2	0	0	2
Replacement	0	5	0	16	3	24
Other	0	1	0	0	38	39
Total	0	137	2	16	45	200

Table 23: HotpotQA D-U → D-V exclusive outcome transitions. Priority is error, reference match, NA, designated replacement match, then other nonmatch; all raw-flag overlaps are zero. Reference uses saved aliases; replacement uses normalized phrase containment.

D-U / D-V	Error	Reference	NA	Replacement	Other	Total
Error	0	0	0	0	0	0
Reference	0	67	0	0	3	70
NA	0	0	0	0	0	0
Replacement	0	1	0	5	1	7
Other	0	0	0	0	23	23
Total	0	68	0	5	27	100

Table 24: MuSiQue D-U → D-V exclusive outcome transitions. Priority is error, reference match, NA, designated replacement match, then other nonmatch; all raw-flag overlaps are zero. Reference uses saved aliases; replacement uses normalized phrase containment.

D-U / D-V	Error	Reference	NA	Replacement	Other	Total
Error	0	0	0	0	0	0
Reference	0	64	0	0	1	65
NA	0	0	2	0	0	2
Replacement	0	4	0	11	2	17
Other	0	1	0	0	15	16
Total	0	69	2	11	18	100

Table 25: All paired reference-match outcomes for D-B relative to D or D-U, stratified by the selected document's post-hoc source category. Stable + / - denotes reference match / nonmatch in both conditions. Source categories do not label sentence truth or identify causes.

Ref.	Source	n	Gained	Lost	Stable +	Stable -
D	all	200	5	11	93	91
D-U	all	200	2	39	96	63
D	target_first	159	3	7	78	71
D-U	target_first	159	2	30	79	48
D	target_second	28	2	2	13	11
D-U	target_second	28	0	4	15	9
D	target_poison	9	0	1	1	7
D-U	target_poison	9	0	4	1	4
D	other_group	4	0	1	1	2
D-U	other_group	4	0	1	1	2

Table 26: HotpotQA paired reference-match outcomes for D-B relative to D or D-U, stratified by the selected document's post-hoc source category. Stable + / - denotes reference match / nonmatch in both conditions. Source categories do not label sentence truth or identify causes.

Ref.	Source	n	Gained	Lost	Stable +	Stable -
D	all	100	3	7	48	42
D-U	all	100	0	19	51	30
D	target_first	84	3	4	41	36
D-U	target_first	84	0	16	44	24
D	target_second	12	0	2	6	4
D-U	target_second	12	0	2	6	4
D	target_poison	2	0	1	0	1
D-U	target_poison	2	0	1	0	1
D	other_group	2	0	0	1	1
D-U	other_group	2	0	0	1	1

Table 27: MuSiQue paired reference-match outcomes for D-B relative to D or D-U, stratified by the selected document's post-hoc source category. Stable + / - denotes reference match / nonmatch in both conditions. Source categories do not label sentence truth or identify causes.

Ref.	Source	n	Gained	Lost	Stable +	Stable -
D	all	100	2	4	45	49
D-U	all	100	2	20	45	33
D	target_first	75	0	3	37	35
D-U	target_first	75	2	14	35	24
D	target_second	16	2	0	7	7
D-U	target_second	16	0	2	9	5
D	target_poison	7	0	0	1	6
D-U	target_poison	7	0	3	1	3
D	other_group	2	0	1	0	1
D-U	other_group	2	0	1	0	1

Table 28: All lexical overlap between selected span S and the visible verified first-hop claim F (mean/median over groups). First/Span coverage divide the unique-token intersection by $|F|/|S|$; Jaccard divides by union size. F in S / S in F count complete normalized phrases appearing contiguously at token boundaries. The frozen answer normalization is used; document titles are excluded. No overlap threshold or outcome-conditioned subgroup is used.

Source	n	First cov.	Span cov.	Jaccard	F in S	S in F
all	200	0.652/0.732	0.346/0.312	0.304/0.288	10	6
target_first	159	0.731/0.800	0.390/0.368	0.350/0.333	10	6
target_second	28	0.329/0.333	0.132/0.102	0.099/0.087	0	0
target_poison	9	0.384/0.429	0.331/0.286	0.218/0.200	0	0
other_group	4	0.363/0.333	0.114/0.115	0.094/0.095	0	0

Table 29: HotpotQA lexical overlap between selected span S and the visible verified first-hop claim F (mean/median over groups). First/Span coverage divide the unique-token intersection by $|F|/|S|$; Jaccard divides by union size. F in S / S in F count complete normalized phrases appearing contiguously at token boundaries. The frozen answer normalization is used; document titles are excluded. No overlap threshold or outcome-conditioned subgroup is used.

Source	n	First cov.	Span cov.	Jaccard	F in S	S in F
all	100	0.675/0.750	0.398/0.366	0.349/0.327	7	6
target_first	84	0.734/0.821	0.437/0.405	0.389/0.377	7	6
target_second	12	0.358/0.314	0.184/0.123	0.129/0.103	0	0
target_poison	2	0.464/0.464	0.315/0.315	0.232/0.232	0	0
other_group	2	0.310/0.310	0.112/0.112	0.088/0.088	0	0

Table 30: MuSiQue lexical overlap between selected span S and the visible verified first-hop claim F (mean/median over groups). First/Span coverage divide the unique-token intersection by $|F|/|S|$; Jaccard divides by union size. F in S / S in F count complete normalized phrases appearing contiguously at token boundaries. The frozen answer normalization is used; document titles are excluded. No overlap threshold or outcome-conditioned subgroup is used.

Source	n	First cov.	Span cov.	Jaccard	F in S	S in F
all	100	0.628/0.667	0.294/0.257	0.259/0.217	3	0
target_first	75	0.727/0.800	0.338/0.312	0.306/0.294	3	0
target_second	16	0.307/0.367	0.092/0.094	0.077/0.081	0	0
target_poison	7	0.361/0.333	0.335/0.286	0.214/0.182	0	0
other_group	2	0.417/0.417	0.115/0.115	0.099/0.099	0	0

F RELATION TO NEIGHBORING RESEARCH INTERFACES

Table 31 compares the aspects needed to place the resource-attribution question. It is a targeted comparison of stated objectives and controls, not a claim that these papers contain no related discussion elsewhere, nor a re-ranking of their experimental performance.

Table 31: Neighboring interfaces and the attribution question addressed here.

Work	Existing objective or control	Relation to this framework
RECOMP (Xu et al., 2023)	Learned evidence compression; outcome-optimized oracle selection	Training supervision differs from test-item oracle access; replacement differs from retained-context quotation
Lost in the Middle (Liu et al., 2023)	Relevant-document position and context length	Motivates utilization sensitivity; position manipulation alone does not estimate verification-message value
Context-faithful prompting (Zhou et al., 2023)	Prompting for contextual faithfulness under conflict and abstention	Changes a fixed model's use of context; contextual obedience and factual recovery have different targets
Sufficient Context / FaithEval (Joren et al., 2025; Ming et al., 2025)	Sufficiency versus utilization; unanswerable, conflicting, and counterfactual contexts	Motivates separating reference agreement, context sufficiency, and appropriate answering
RobustRAG / TrustRAG (Xiang et al., 2024; Zhou et al., 2025)	Response isolation/aggregation or filtering/self-assessment against corruption	Defense mechanisms with their own assumptions; our controlled annotations and displays do not establish a superiority ranking
This work	Audited selector access and matched-span status display	Reports controlled intervention effects and explicitly separates them from unestimated information advantages

The classical decision principles establish how ideal risks compare across disclosed observations. Our domain-specific construction additionally records the selector's access to annotations, the span and source binding exposed to the generator, and the distinction between acquiring a variable verification message and displaying a fixed positive status. These interfaces determine the report in Table 2. Its practical output is a qualified comparison, not an estimator of latent information terms or a new deployed defense.

G CANDIDATE REVIEW AND SELECTION DETAILS

Table 32 compares decisions on identical direct candidates. Under B1, ordinary review changes Qwen3 matches $106 \rightarrow 98$ and Qwen3.8 matches $102 \rightarrow 53$, with no newly matched answers. Qwen3 loses seven matches to abstention and one to another nonmatching answer; all 49 Qwen3.8 losses are to abstention. Qwen3 coverage changes $95.0\% \rightarrow 87.5\%$ while answered-reference quality changes $55.8\% \rightarrow 56.0\%$. Qwen3.8 coverage changes $81.5\% \rightarrow 40.0\%$ while quality changes $62.6\% \rightarrow 66.3\%$. Thus the latter model's fewer nonmatching answers primarily reflect a more restrictive answering policy. Withdrawing an insufficiently supported candidate can comply with the answering instruction even when its reference value matches; the coverage loss alone does not establish a reliability failure. Benign B0 review also yields no gains and loses three/one matches. This is a result for these prompts and models; the exploratory pilot contains some review gains.

Sufficiency gating and output-budget completion. The original gate truncates on 65 Qwen3 and 182 Qwen3.8 judgments. All 247 declared retries finish within the longer budget, with no remaining interface failures, while 553 original judgments are reused unchanged. This separates the gate's decision behavior from its initial output-budget failure. Under B1, the completed gate retains only 18/15 answers, of which 4/5 match the reference: coverage is $9.0\%/7.5\%$, and answered-reference quality is $22.2\%/33.3\%$. At that same coverage, random retention of the valid direct candidates gives expected match counts of 10.04/9.39 (Table 12). At these descriptive operating points, reference selection does not exceed the random-retention expectation. That comparison

Table 32: Same-candidate decisions with ordinary successor content (200 groups per row).

Model	Input	Method	EM count	Coverage%	Quality%	Fail	Tokens/query
Qwen3-32B	B0	Direct	151	98.0	77.0	0	454
Qwen3-32B	B0	Review	148	97.0	76.3	0	962
Qwen3-32B	B0	SC-768	107	70.0	76.4	25	2131
Qwen3-32B	B0	SC-completed	120	78.5	76.4	0	2364
Qwen3-32B	B1	Direct	106	95.0	55.8	0	370
Qwen3-32B	B1	Review	98	87.5	56.0	0	795
Qwen3-32B	B1	SC-768	2	5.5	18.2	40	1983
Qwen3-32B	B1	SC-completed	4	9.0	22.2	0	2339
Qwen3.8-27B	B0	Direct	166	98.5	84.3	0	461
Qwen3.8-27B	B0	Review	165	98.0	84.2	0	976
Qwen3.8-27B	B0	SC-768	107	61.5	87.0	71	2186
Qwen3.8-27B	B0	SC-completed	156	90.5	86.2	0	2910
Qwen3.8-27B	B1	Direct	102	81.5	62.6	0	377
Qwen3.8-27B	B1	Review	53	40.0	66.3	0	805
Qwen3.8-27B	B1	SC-768	0	0.5	0.0	111	2077
Qwen3.8-27B	B1	SC-completed	5	7.5	33.3	0	3244

Tokens/query expands shared candidate generation plus decision costs as if evaluating that method alone. SC-completed includes both its initial gate and length retry; it is not a fresh 4096-token gate on every query. Quality is reference-match fraction among valid non-NA answers.

evaluates reference agreement, not whether answering is warranted by the context, and does not establish a general ranking of sufficient-context methods.

Without competition, the completed gate retains 120/156 matches at $78.5\%/90.5\%$ coverage. Its difference from matched-coverage random retention is $-0.95/+3.48$ matches. The additional calls consume 495881 tokens; the full candidate-plus-gating costs appear in Table 32. By construction the gate cannot create a match absent from its candidate. Its intended role is answer selection using predicted context sufficiency. We therefore report matched-answer coverage, reference quality and cost separately, without interpreting every rejected reference match as an inappropriate abstention.

H Connection to Graph Recovery

The initial LightRAG (Guo et al., 2024) study has 24 groups and six query types per group. Compositional correctness across cleaning, first-hop verification, and fully trusted support is $12 \to 13 \to 19$ for Qwen and $7 \to 5 \to 10$ for Mistral. The full-support arm changes both content and trust, motivating the framework's separate interfaces. A 16-group pilot further finds that ordinary successor evidence helps without competition, while review under competition mainly suppresses answers. These exploratory results use semantic scoring and different inputs; we do not pool them with the reference-based confirmation. Appendices I–P retain their full results and sensitivity analyses.

I EXPLORATORY STUDIES: SETUP

I.1 INFORMATION RESOURCES

A verification record consists of a qualified claim and its source excerpt. The claim includes subject, relation, object, and necessary time/entity constraints. Provenance records where the statement came from; only explicitly verified claims receive verification status.

Table 33: Information interfaces and their interpretation.

Resource	Information supplied	Interpretation
Error feedback / cleaning	Query and an incorrect answer confirmed by the strict interface, for tracing and cleaning	No correct answer or trusted full source is supplied; the interface may return unknown.
R1: atomic verification	A verified claim and supporting excerpt for each preselected first-hop query	All 24 predetermined facts are provided, not just previously incorrect queries; surrounding text is not automatically trusted.
R2: trusted full support	The corresponding two complete original supporting documents, explicitly treated as trusted	Adds successor facts and stronger trusted information; a different resource condition.

R1/R2 are controlled benchmark-derived interfaces, not deployed fact-checkers. Scope policies are compared within a resource level; added information is disclosed across levels.

I.2 STAGES AND INDEPENDENT UNITS

Table 34: Experimental stages and limits of inference.

Stage	Sample and role	Inference limit
S1: fact scope	24 groups, 12 each from HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022); 6 query types per group; actual LightRAG construction, deletion, retrieval, then two generators	One complete graph-building pipeline; 144 queries are not 144 independent facts; groups may share general entities or themes.
S2: completion diagnostics	Reuses S1's 24 groups; ordinary original text retrieved from the remaining corpus using the query, bridge information, or verified object; same-candidate review	Ablations on observed data, not new independent confirmation; mechanism diagnostics only.
S3: evidence control	Another 16 groups, 8 per dataset; first-hop and composition queries; successor availability crossed with three pack types	Oracle-selected successor sentences; fixed natural-language packs, without new retrieval or graph construction.

S1's primary generator is Qwen3-32B-AWQ (Qwen Team, 2025); Mistral-7B-Instruct-v0.3 (Mistral AI, 2024) reanswers the same Qwen graph contexts. Scale and quantization also differ, so differences cannot be attributed solely to model family. Stage protocols and input hashes are retained. S1 contains 3168 logical records ($24 \times 6 \times 11 \times 2$); the three displayed resource conditions comprise 864. S3 contains 768 ($16 \times 2 \times 6 \times 2 \times 2$); its composition slice contains 384. Logical records are neither independent calls nor independent fact groups. S1 and S3 are not pooled. Appendix P specifies counting and scoring; Appendices L–M retain the complete displayed conditions.

I.3 EVALUATION

Frozen scoring retains alias exact match (EM), token F1, strict NA, pipeline failure, and designated incorrect-phrase matches. Supplemental automated semantic review of non-exact answers distinguishes fully correct, incorrect, abstention, partial, uncertain, and failure; it is not independent human gold labeling. Same-model, same-query comparisons count non-correct-to-correct gains, correct-to-incorrect/abstaining/failing losses, and state transitions. Fact groups are the analysis unit;

paraphrases, order replays, and cached reuse do not increase sample size. Figures show observed counts, without transferring EM intervals to semantic accuracy.

Independent reaggregation and figure/table review are complete. The Venice-with-incorrect-Corelli-explanation and correct-century-with-fabricated-verification cases were human-adjudicated as partial under whole-response semantics. Other independent gold labels for evidence sufficiency, trusted conflict resolvability, and appropriate actions remain incomplete. We do not combine reference correctness, visible support, and abstention into a validated safety score. Section J.4 supplements the frozen labels only for partial/uncertain records in specified slices; it neither overwrites them nor rejudges incorrect answers or interface failures.

J Exploratory Studies: Results

J.1 Verification resources and recovery scope

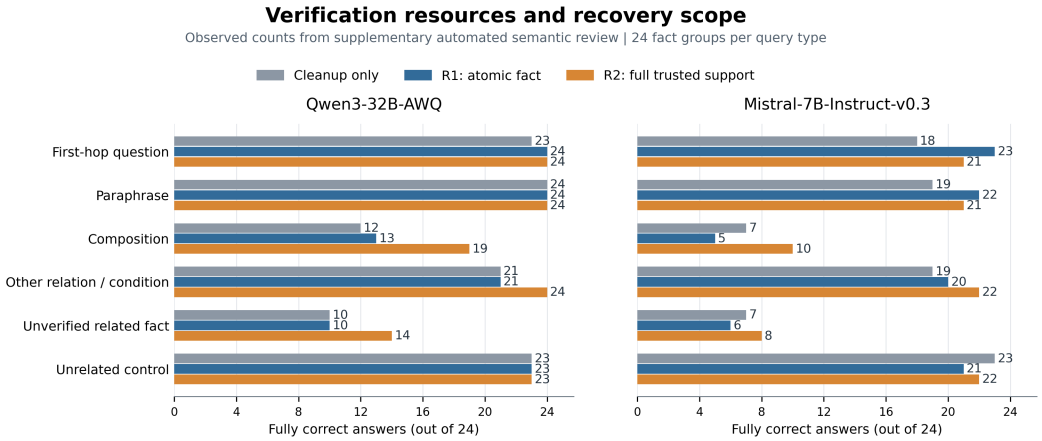

Figure 6: Fully correct semantic answers by query scope in S1, each category containing 24 fact groups. Cleaning, R1-all, and R2-all are shown by generator. R2 adds trusted content and is not an equal-information algorithm comparison. Supplemental automated judgments retain failures in denominators; counts have no error bars. Full EM/F1 and states appear in Appendix L.

Composition correctness is $12 \rightarrow 13 \rightarrow 19$ for Qwen and $7 \rightarrow 5 \rightarrow 10$ for Mistral. Only 1 fact group gains under R1 in both generators under whole-response semantics; that exact count is scoring-sensitive (Section J.4). Full documents contain additional successor facts, so higher R2 correctness does not establish stronger propagation of atomic correction. Similarity-gating diagnostics likewise show that similar questions in the same group can request different relations; entity matching is not final-target equivalence.

Input review found that 27 S1 non-composition queries selected a top-1 verification record from another construction group: 13 unrelated controls, 12 unverified related queries, and 2 different-relation queries. All 24 composition queries selected their own group's record. Out-of-scope outcomes therefore combine record selection and use rather than represent purely "no new verification"; they cannot uniformly be attributed to trust spillover from the corresponding fact. Cross-group selection is not automatically a retrieval error.

J.2 Successor visibility and competing claims

Without competition, E0→E1 raises Qwen/Mistral composition correctness from 0/1 to 14/10. Separate competition yields 5/3 and same-source mixing 4/4. Completing ordinary content differs from obtaining trusted information that distinguishes truth from falsehood. Role, unit, and relation errors remain under complete benign input: visible evidence is not sufficient for correct execution.

Mixed-versus-separate directions vary across models and review conditions; "same-source mixing is more dangerous" is not a central claim. The sales case compares compatible quantitative lower

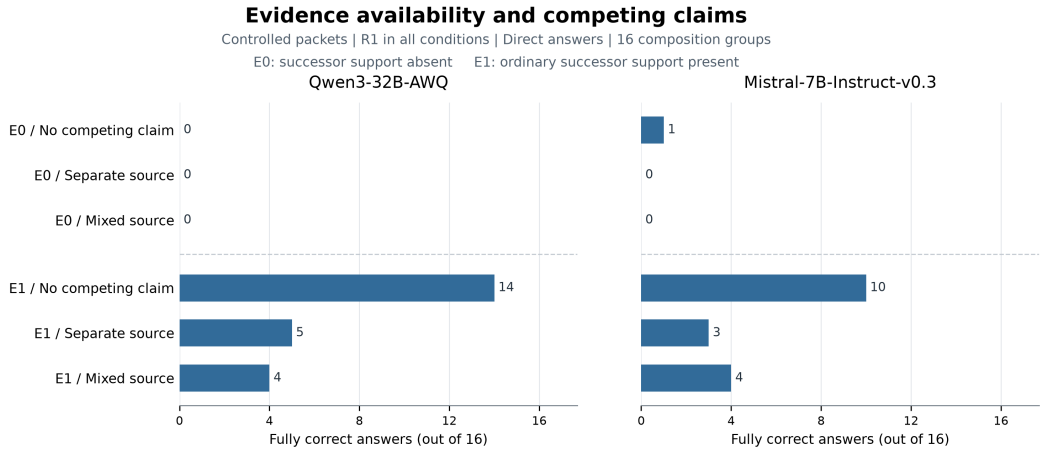

Figure 7: S3 direct-answer semantic correctness across six evidence conditions, with 16 groups per condition. E0 lacks genuine successor support; E1 supplies an ordinary original successor sentence. None/separate/mixed denote absent competition, a separate simulated source, and same-source competing claims. Pack counts are fixed, but text lengths are not exactly equal. Oracle-provided E1 support is not retrieval recall. Failures remain in denominators; all review conditions appear in Appendix M.

bounds. We retain all 16 groups and the existing post-hoc analysis excluding that case; it is not evidence of a strict logical contradiction. Co-production, jointly named artists, and multivalued attributes still require case-specific adjudication. Removing the sales group does not establish that the remaining 15 cases are strictly mutually exclusive.

J.3 Correct recovery and abstention

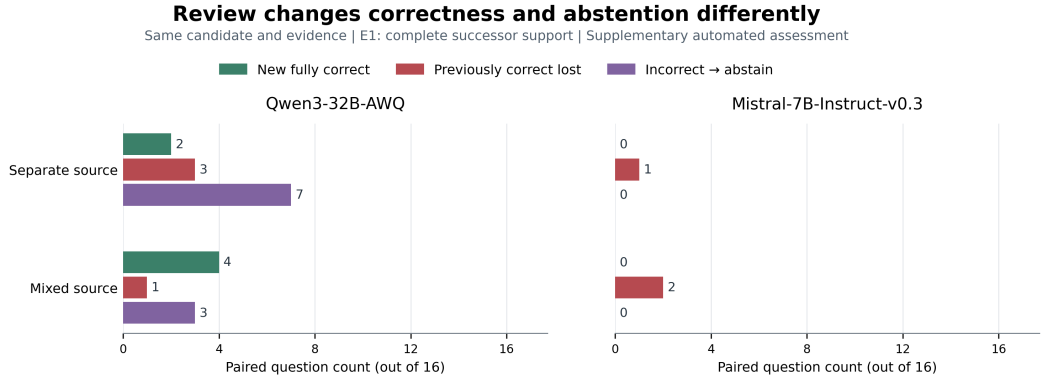

Figure 8: S3 review versus direct answering with complete evidence and separate/mixed competition. Each model–condition pair contains the same 16 queries. Gains, losses, and incorrect-to-abstention transitions are paired by ID, not inferred from aggregate subtraction. Table 35 separates other states; Appendix N gives full matrices. Review uses extra computation and is not an equal-compute comparison.

Qwen E1-separate gains 2 correct answers and loses 3, changing correctness $5 \rightarrow 4$; the error decrease $11 \rightarrow 2$ mainly accompanies more abstention. E1-mixed gains 4 and loses 1, changing correctness $4 \rightarrow 7$, a local benefit. Mistral has no corresponding net recovery gain under either competing condition. Its benign correctness remains $10 \rightarrow 7$, with the three losses decomposed in Table 36.

Table 35: Paired review effects under complete evidence and competing claims.

Model	Condition	C before→after	Gain	Loss	Loss W/A/P/U/E	W→A	Net
Qwen	e1_separate	5→4	2	3	0/3/0/0/0	7	-1
Qwen	e1_mixed	4→7	4	1	0/1/0/0/0	3	3
Mistral	e1_separate	3→2	0	1	0/1/0/0/0	0	-1
Mistral	e1_mixed	4→2	0	2	1/1/0/0/0	0	-2

Each model–condition pair contains 16 paired records. Loss includes any fully-correct-to-other-state transition. W/A/P/U/E denote incorrect, abstention, partial, uncertain, and pipeline failure. None of the correct-answer losses in these four rows is an interface failure.

Table 36: The three correct-answer losses for Mistral with benign complete evidence.

Query ID	Correct candidate / target	Review endpoint	Transition
`q-6ce329ba3e907a0d53af5025`	Awards: 6 Silver Slugger Awards	Accept did not preserve candidate wording; contract failure	Correct → interface failure
`q-02cfbc3a70d325c8eb44320d`	Song title: The Land of Enchantment	Accept removed candidate quotation marks; contract failure	Correct → interface failure
`q-abf7a46aa518e7438c88575c`	Sport: candidate included competitive swimming	Valid final output only "Sports film", not the requested sport	Correct → incorrect

Thus, 2 losses are interface failures and 1 is an incorrect answer. Interface usability loss and content degradation must be separated; these cases neither establish general reasoning degradation nor show that retrying caused the error. The additional interface failure for `q-15511fdcc0bf28868d10fc6d` starts from an incorrect answer and is not among these correct-answer losses.

Abstention is not uniformly a failure. When ordinary sources disagree and trusted records cover only the first hop, the generator may lack grounds to choose the true successor fact. Evaluation should distinguish factual recovery from uncertainty management.

J.4 Whole-response and final-answer sensitivity

Whole-response semantics remains primary, including the two human-adjudicated partial cases. We rescore only S1 composition under cleaned/R1-all/R2-all and S3 composition direct answers under E0-none/E1-none/E1-separate/E1-mixed, for both generators. Denominators remain 24 or 16, retaining failures. Only a frozen partial/uncertain record whose finally asserted answer value matches a reference or frozen alias becomes correct in the final-answer view; all other states remain unchanged. Matching uses the frozen lowercase, ASCII-punctuation, article, and whitespace normalization, followed by whole-string equality: no substrings, synonym expansion, date completion, or selecting a true alternative from an "or" list. This is not an EM or semantic rerun over all outputs. Appendix P specifies the rule.

Table 37: Whole-output and final-answer scoring sensitivity.

Stage / Model	Whole output (W)	Final answer (F)	n per cell
S1 / Qwen	12 → 13 → 19	13 → 13 → 19	24
S1 / Mistral	7 → 5 → 10	8 → 5 → 10	24
S3 / Qwen	0 → 14 → 5 → 4	1 → 14 → 5 → 4	16
S3 / Mistral	1 → 10 → 3 → 4	1 → 10 → 3 → 4	16

S1 condition order: cleaned → R1-all → R2-all. S3 condition order: E0-none → E1-none → E1-separate → E1-mixed.

Only 3 partial records change: Qwen cleaned's terminal Delaware River, Mistral cleaned's Venice, and Qwen E0-none's explicit "Answer: 15th century". They respectively accompany an incorrect intermediate watercourse, incorrect Corelli explanation, and fabricated V4 verification marker. Final-

value matches do not remove whole-response defects. The remaining 6 partial/uncertain cases retain their labels; Appendix O reports all 14 cells.

The single shared R1 composition-gain group exists only under whole-response scoring. Final-answer scoring already counts Mistral's original Venice as correct, leaving 0 shared-gain groups. Pairing against cleaned, Qwen's R1 gains/losses change from 1/0 to 1/1; Mistral's from 1/3 to 0/3. Neither that exact shared count nor Qwen's small net increase is robust across scoring targets. Neither view shows consistent net composition improvement across generators; higher R2 correctness still involves stronger resources. Qwen E0 changing $0 \rightarrow 1$ demonstrates an answer-value match, not visible or verified successor support; E1 cells are unchanged. S3 review outputs were not rescored, so Figure 8 retains frozen transitions and is not claimed to have comprehensive dual-criterion validation.

K Conditional Information Boundary

Let the complete observation x contain the query, visible evidence, verification records, all trusted background, and a fixed generator. Suppose two admissible environments yield the same x, satisfy all verified facts, but have mutually exclusive correct final answers y_0, y_1. An answerer depending only on x outputs a single answer or abstains. If its answer probability is c and e_i is its probability of answering incorrectly in environment i, then

$$e_0 + e_1 = 2c - P(A = y_0 \mid x) - P(A = y_1 \mid x) \geq c. \tag{4}$$

In at least one admissible environment, the incorrect-answer probability is at least $c/2$; requiring zero error in both environments at this observation forces abstention. The argument uses only identical output distributions and mutually exclusive correct answers. Distinguishing trusted information changes the observation and escapes that original indistinguishability premise.

This is a task-specific rendering of a classical indistinguishability argument, *not a new theorem*. It applies only under its assumptions, does not imply a dataset-wide error lower bound of 50%, and does not deny the statistical value of unverified sources or review's ability to repair execution errors. S3 did not construct and test identical-observation twin environments: its evidence is compatible with, not a validation of, the proposition. Risk and coverage relate to selective classification (El-Yaniv & Wiener, 2010), but its guarantees cannot be transplanted to RAG without checking their assumptions.

L VERIFICATION RESOURCES AND QUERY SCOPE

Each row contains the same S1 fact groups for a given query category. Categories must not be pooled as independent facts. Both generators share graph contexts constructed with Qwen. R2 supplies additional content and trusted information, so this is not an equal-resource algorithm comparison.

Table 38: Verification resources by query scope: Qwen.

Method	Scope	n	C	W	A	P	U	E	EM	F1	NA
cleaned	First hop	24	23	0	0	1	0	0	22	0.9457	0
cleaned	Paraphrase	24	24	0	0	0	0	0	24	1.0	0
cleaned	Composition	24	12	10	0	1	0	1	9	0.4754	0
cleaned	Other relation	24	21	1	2	0	0	0	20	0.8611	2
cleaned	Unverified	24	10	14	0	0	0	0	10	0.4167	0
cleaned	Control	24	23	0	1	0	0	0	21	0.9236	1
r1_all	First hop	24	24	0	0	0	0	0	23	0.9722	0
r1_all	Paraphrase	24	24	0	0	0	0	0	24	1.0	0
r1_all	Composition	24	13	9	2	0	0	0	12	0.5536	1
r1_all	Other relation	24	21	1	2	0	0	0	19	0.8458	2
r1_all	Unverified	24	10	14	0	0	0	0	10	0.4167	0
r1_all	Control	24	23	0	1	0	0	0	20	0.904	1
r2_all	First hop	24	24	0	0	0	0	0	22	0.9567	0
r2_all	Paraphrase	24	24	0	0	0	0	0	23	0.9712	0
r2_all	Composition	24	19	5	0	0	0	0	17	0.7729	0
r2_all	Other relation	24	24	0	0	0	0	0	22	0.9708	0
r2_all	Unverified	24	14	10	0	0	0	0	14	0.5833	0
r2_all	Control	24	23	0	1	0	0	0	21	0.9179	1

C: fully correct; W: incorrect; A: abstention; P: partial; U: uncertain; E: pipeline failure. NA is the frozen strict-NA count. All states remain in the denominator; semantic labels are supplemental automated judgments, not independently validated human gold labels.

Table 39: Verification resources by query scope: Mistral.

Method	Scope	n	C	W	A	P	U	E	EM	F1	NA
cleaned	First hop	24	18	5	1	0	0	0	14	0.7005	0
cleaned	Paraphrase	24	19	4	0	1	0	0	16	0.7611	0
cleaned	Composition	24	7	14	1	1	0	1	6	0.3279	0
cleaned	Other relation	24	19	3	1	1	0	0	17	0.7858	0
cleaned	Unverified	24	7	16	0	0	1	0	3	0.2397	0
cleaned	Control	24	23	0	1	0	0	0	12	0.7385	0
r1_all	First hop	24	23	1	0	0	0	0	18	0.8445	0
r1_all	Paraphrase	24	22	2	0	0	0	0	20	0.8746	0
r1_all	Composition	24	5	18	1	0	0	0	4	0.3308	0
r1_all	Other relation	24	20	2	2	0	0	0	17	0.765	0
r1_all	Unverified	24	6	18	0	0	0	0	2	0.2161	0
r1_all	Control	24	21	1	1	0	1	0	16	0.7793	0
r2_all	First hop	24	21	3	0	0	0	0	15	0.7616	0
r2_all	Paraphrase	24	21	3	0	0	0	0	18	0.811	0
r2_all	Composition	24	10	13	0	1	0	0	5	0.4289	0
r2_all	Other relation	24	22	1	1	0	0	0	20	0.8614	0
r2_all	Unverified	24	8	16	0	0	0	0	5	0.2953	0
r2_all	Control	24	22	0	1	0	1	0	19	0.8589	0

C: fully correct; W: incorrect; A: abstention; P: partial; U: uncertain; E: pipeline failure. NA is the frozen strict-NA count. All states remain in the denominator; semantic labels are supplemental automated judgments, not independently validated human gold labels.

M SUCCESSOR EVIDENCE AND SOURCE-PACK CONDITIONS

S3 uses a separate set of fact groups and composition queries. E0 lacks genuine successor support; E1 contains ordinary genuine successor support. The source-pack conditions are none, separate, and mixed. Support is oracle-selected; this experiment does not measure retrieval performance. The full denominator includes the sales lower-bound case, whose claims are not strictly mutually exclusive.

Table 40: Successor evidence and competing claims: Qwen.

Condition	Method	n	C	W	A	P	U	E	EM	F1	NA	Hit/n
e0_none	answer	16	0	4	10	1	1	0	0	0.0581	10	0/16
e0_separate	answer	16	0	15	1	0	0	0	0	0.2167	1	9/16
e0_mixed	answer	16	0	15	1	0	0	0	0	0.2167	1	10/16
e1_none	answer	16	14	2	0	0	0	0	13	0.8482	0	0/16
e1_separate	answer	16	5	11	0	0	0	0	3	0.417	0	7/16
e1_mixed	answer	16	4	12	0	0	0	0	3	0.3482	0	7/16
e0_none	review	16	0	0	15	0	0	1	0	0.0	15	0/16
e0_separate	review	16	0	13	2	0	0	1	0	0.2167	2	9/16
e0_mixed	review	16	0	13	1	0	0	2	0	0.2167	1	9/16
e1_none	review	16	13	1	1	0	0	1	12	0.7857	1	0/16
e1_separate	review	16	4	2	10	0	0	0	3	0.2649	10	1/16
e1_mixed	review	16	7	5	4	0	0	0	6	0.442	4	2/16

C: fully correct; W: incorrect; A: abstention; P: partial; U: uncertain; E: pipeline failure. NA is the frozen strict-NA count. All states remain in the denominator; semantic labels are supplemental automated judgments, not independently validated human gold labels. Hit/n: frozen attack-phrase matches / annotated denominator.

Table 41: Successor evidence and competing claims: Mistral.

Condition	Method	n	C	W	A	P	U	E	EM	F1	NA	Hit/n
e0_none	answer	16	1	9	6	0	0	0	1	0.1116	0	0/16
e0_separate	answer	16	0	15	1	0	0	0	0	0.1919	0	8/16
e0_mixed	answer	16	0	15	1	0	0	0	0	0.2033	0	7/16
e1_none	answer	16	10	6	0	0	0	0	8	0.6181	0	0/16
e1_separate	answer	16	3	10	0	1	2	0	3	0.3317	0	6/16
e1_mixed	answer	16	4	11	0	0	1	0	4	0.3429	0	4/16
e0_none	review	16	1	6	5	0	0	4	1	0.0938	5	0/16
e0_separate	review	16	0	14	1	0	0	1	0	0.1646	0	6/16
e0_mixed	review	16	0	13	1	0	0	2	0	0.172	0	6/16
e1_none	review	16	7	6	0	0	0	3	7	0.5	0	0/16
e1_separate	review	16	2	11	1	1	1	0	2	0.2812	1	6/16
e1_mixed	review	16	2	11	1	0	1	1	2	0.2396	1	4/16

C: fully correct; W: incorrect; A: abstention; P: partial; U: uncertain; E: pipeline failure. NA is the frozen strict-NA count. All states remain in the denominator; semantic labels are supplemental automated judgments, not independently validated human gold labels. Hit/n: frozen attack-phrase matches / annotated denominator.

N Complete Review-Transition Matrices

Rows are states before review and columns are states after review, paired by query ID. Each matrix reuses the corresponding E1 records in Appendix M; it adds no fact groups. A dash denotes no observed transition. The displayed nonzero counts are exactly those in the frozen transition lists.

Table 42: Review transitions: Qwen, e1_separate.

Before / After	C	W	A	P	U	E
C	2	—	3	—	—	—
W	2	2	7	—	—	—
A	—	—	—	—	—	—
P	—	—	—	—	—	—
U	—	—	—	—	—	—
E	—	—	—	—	—	—

C: fully correct; W: incorrect; A: abstention; P: partial; U: uncertain; E: pipeline failure. NA is the frozen strict-NA count. All states remain in the denominator; semantic labels are supplemental automated judgments, not independently validated human gold labels.

Table 43: Review transitions: Qwen, e1_mixed.

Before / After	C	W	A	P	U	E
C	3	—	1	—	—	—
W	4	5	3	—	—	—
A	—	—	—	—	—	—
P	—	—	—	—	—	—
U	—	—	—	—	—	—
E	—	—	—	—	—	—

C: fully correct; W: incorrect; A: abstention; P: partial; U: uncertain; E: pipeline failure. NA is the frozen strict-NA count. All states remain in the denominator; semantic labels are supplemental automated judgments, not independently validated human gold labels.

Table 44: Review transitions: Mistral, e1_separate.

Before / After	C	W	A	P	U	E
C	2	—	1	—	—	—
W	—	10	—	—	—	—
A	—	—	—	—	—	—
P	—	—	—	1	—	—
U	—	1	—	—	1	—
E	—	—	—	—	—	—

C: fully correct; W: incorrect; A: abstention; P: partial; U: uncertain; E: pipeline failure. NA is the frozen strict-NA count. All states remain in the denominator; semantic labels are supplemental automated judgments, not independently validated human gold labels.

Table 45: Review transitions: Mistral, e1_mixed.

Before / After	C	W	A	P	U	E
C	2	1	1	—	—	—
W	—	10	—	—	—	1
A	—	—	—	—	—	—
P	—	—	—	—	—	—
U	—	—	—	—	1	—
E	—	—	—	—	—	—

C: fully correct; W: incorrect; A: abstention; P: partial; U: uncertain; E: pipeline failure. NA is the frozen strict-NA count. All states remain in the denominator; semantic labels are supplemental automated judgments, not independently validated human gold labels.

O SCORING SENSITIVITY

W is the frozen whole-output semantic criterion. F changes only partial or uncertain records in the specified slices when the extracted final answer matches a reference or frozen alias under the frozen normalization rules. All other labels and denominators remain unchanged.

Table 46: All scoring-sensitivity cells.

Stage	Model	Condition	Method	n	W	F	F−W
S1	Qwen	cleaned	cleaned	24	12	13	1
S1	Qwen	r1_all	r1_all	24	13	13	0
S1	Qwen	r2_all	r2_all	24	19	19	0
S1	Mistral	cleaned	cleaned	24	7	8	1
S1	Mistral	r1_all	r1_all	24	5	5	0
S1	Mistral	r2_all	r2_all	24	10	10	0
S3	Qwen	e0_none	answer	16	0	1	1
S3	Qwen	e1_none	answer	16	14	14	0
S3	Qwen	e1_separate	answer	16	5	5	0
S3	Qwen	e1_mixed	answer	16	4	4	0
S3	Mistral	e0_none	answer	16	1	1	0
S3	Mistral	e1_none	answer	16	10	10	0
S3	Mistral	e1_separate	answer	16	3	3	0
S3	Mistral	e1_mixed	answer	16	4	4	0

S1 uses composition queries under cleaned, R1-all, and R2-all; S3 uses composition-query direct answers under E0-none, E1-none, E1-separate, and E1-mixed. Review outputs are not reclassified.

Table 47: Paired sensitivity of R1 relative to cleaning for S1 composition queries.

Model	Criterion	Paired n	Cleaned C	R1 C	Gain	Loss	Net
Qwen	W (whole output)	24	12	13	1	0	1
Qwen	F (final answer)	24	13	13	1	1	0
Mistral	W (whole output)	24	7	5	1	3	-2
Mistral	F (final answer)	24	8	5	0	3	-3

Gain is a non-correct-to-correct transition; loss is correct-to-any-other-state. The intersection of fact groups gaining correctness for both generators is 1 under W and 0 under F; it is not the sum of their gain counts.

P Record Accounting and Evaluation Protocol

P.1 Construction and information boundaries

S1 uses 24 groups from HotpotQA and MuSiQue, 12 from each. A natural two-hop item provides a first-hop verification query and a composition query whose target requires an additional fact. The six categories are the first-hop query, its paraphrase, a new composition query, another relation or condition for the same entity, an unverified related fact, and an unrelated control. Groups were fixed before predictions. R1 provides a preselected qualified atomic claim and supporting excerpt for every first-hop query, not just those initially judged wrong. Only the specified claim is verified. R2 supplies the corresponding two full original support documents as trusted resources and includes extra successor information. Actual selection of a verification record remains part of the retrieval policy.

S2 reuses the same 24 groups for ordinary-text completion and same-candidate review; it is a development diagnostic, not an independent sample. S3 uses a separate, source-disjoint 16 groups, 8 from each dataset, with first-hop and composition queries. Its fixed packs vary successor visibility (E0/E1) and competition (none/separate/mixed). E1 is an oracle-selected ordinary successor sentence, not a retrieved or externally verified successor fact. The sales case has compatible lower bounds; keeping the full denominator does not certify strict exclusivity of all competing statements. Other multivalued attributes also require case-specific qualification.

P.2 Logical records versus displayed slices

Table 48: Full record counts, displayed slices, and independent units.

Stage	Full logical records	Displayed slice	Fact groups
S1	$3168 = 24 \times 6 \times 11 \times 2$	$864 = 24 \times 6 \times 3 \times 2$ for cleaned/R1-all/R2-all	24
S3	$768 = 16 \times 2 \times 6 \times 2 \times 2$	$384 = 16 \times 1 \times 6 \times 2 \times 2$ for composition	16

S1's factors are groups, query categories, conditions, and generators. S3's are groups, query categories, evidence conditions, methods, and generators. Logical records include cached reuse and are not model-call counts. The displayed slice must not be mistaken for the full record set, and the two stages must not be pooled into one estimate. Review transitions reuse direct-answer queries and add no groups. S3 first-hop controls are 16/16 semantically correct in each condition and are excluded from the composition tables. The 16 final S3 failures all occur in review. The original 16-group results retain the sales case; the existing 15-group post-hoc sensitivity does not replace the primary denominator.

P.3 Frozen scoring and supplemental semantics

EM matches the normalized prediction against a reference or frozen alias. Normalization lowercases, deletes ASCII punctuation, removes the English articles `a`, `an`, and `the`, and collapses whitespace. Token F1, strict NA, pipeline failure, and designated incorrect-phrase matching retain their frozen definitions. Strict NA can differ from semantic abstention. Fully correct, incorrect, abstention, partial, uncertain, and failure all remain in the same denominator. Supplemental semantic judgments are automated assessments rather than a complete independent human gold standard. Two boundary outputs were adjudicated as partial under the primary whole-response criterion; that does not resolve every other case.

Gains pair non-correct outputs with correct outputs for the same generator and query; losses pair correct outputs with any other state. Incorrect-to-abstention is not total abstention: Qwen separate has 7 incorrect and 3 correct answers becoming abstentions. Interface failures must remain distinct from content errors and cannot be dropped from denominators. EM/F1, attack-phrase hits, and semantic correctness answer different questions and do not replace each other. Descriptive figures have no semantic confidence intervals; absence of error bars is not evidence of population-level cer-

tainty. Any inference requires fact-group-level analysis of a specified paired estimand, not treating method, model, or paraphrase records as independent samples.

P.4 Final-answer sensitivity rule

Only the specified partial/uncertain records are eligible: S1 composition under cleaned/R1-all/R2-all for both generators, and S3 composition direct answers under E0-none/E1-none/E1-separate/E1-mixed for both generators. These slices contain 144 and 128 records, respectively; their 272 records are not independent samples. Every other frozen state is retained, including incorrect, abstaining, and failing outputs. Eligible records become correct only if the finally asserted answer value equals a reference or frozen alias after the same normalization. This is whole-string equality, not arbitrary substring matching, semantic synonym expansion, date completion, or choosing a true alternative from an unresolved "or" list.

Of 9 eligible records, 3 change and 6 do not. The changed values are Delaware River in Qwen cleaned, Venice in Mistral cleaned, and 15th century in Qwen E0-none. Wrong intermediate entities, wrong explanations, and fabricated verification status remain defects under whole-response scoring. The scoring-policy comparison does not overwrite primary labels. R1's common-gain intersection across the generators is 1 under whole-response scoring and 0 under final-answer scoring; it is not the sum of individual gains. Review outputs are outside this rescoring scope, so the frozen review matrices do not claim comprehensive dual-criterion robustness.