INTERVENERAG: DIAGNOSING WHERE GRAPHRAG ANSWERS CAN BE RECOVERED

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) helps large language models answer knowledge-intensive questions by supplying external evidence. GraphRAG organizes this evidence into graphs for questions requiring facts from multiple sources. Incorrect answers may arise from missing graph information, retrieval failures, or ineffective evidence use, but component metrics do not directly show whether a stage repair would improve the answer. We introduce **InterveneRAG**, a diagnostic framework measuring *recoverability*: the paired change in answer quality under a specified stage intervention. Annotation-guided construction and retrieval replacements and evidence-ordering probes quantify remaining repair potential with downstream operators fixed. Individual and joint interventions, together with graph-size and context-budget controls, examine how gains depend on stage interactions and evidence selection and presentation. Across two multi-hop question-answering datasets and three generator families, annotation-guided selection within existing automatic graphs improves answer F1 by 0.206–0.339, despite those graphs containing about 95% of annotated supporting facts. Extensions with two generators yield F1 gains of 0.293–0.305 on an equal-hop MuSiQue cohort and 0.086–0.117 through selection restricted to existing passages in bounded shared HippoRAG indices. Construction–retrieval interactions vary across comparison protocols; on 2WikiMultiHopQA, a new factorial protocol combining common budget ceilings with a revised retrieval intervention yields interaction estimates whose confidence intervals cross zero. These findings reveal substantial selection headroom within existing graphs and indices while showing that stage interactions require protocol-specific interpretation. The measured gains characterize annotation-guided repair potential rather than guaranteed deployable improvements. Code: `https://anonymous.4open.science/r/InterveneRAG-1E85/`

1 INTRODUCTION

Retrieval-augmented generation (RAG) supplies external evidence for knowledge-intensive questions (Lewis et al., 2020). Multi-hop questions require connecting facts across sources. Graph-based RAG organizes entities and relations to support this process (Jiménez Gutiérrez et al., 2024; Guo et al., 2024). In our pipeline, construction extracts a provenance-linked graph, retrieval selects query-relevant evidence, and serialization presents it to the generator. Answer quality depends on both the represented information and how it reaches the model.

A necessary fact may be absent from the graph, missed by retrieval, or retrieved but ineffectively used. Existing evaluation frameworks measure relevance, coverage, faithfulness, and answer quality (Es et al., 2024; Ru et al., 2024). These characterize an execution but do not directly answer: *how would the final answer change under a specified stage repair?* Increasing coverage can recover a missing fact yet leave the answer unchanged if the generator cannot combine the evidence. Component quality and repair response therefore serve complementary purposes.

Stage replacement can change graph size, candidate evidence, and context amount, composition, or order. Construction and retrieval may also recover overlapping improvements, requiring explicit interventions and matched graph/context controls.

We introduce INTERVENERAG (Figure 1), defining *recoverability* as the paired answer-quality change under a specified intervention. Construction and retrieval replacements use annotations;

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.

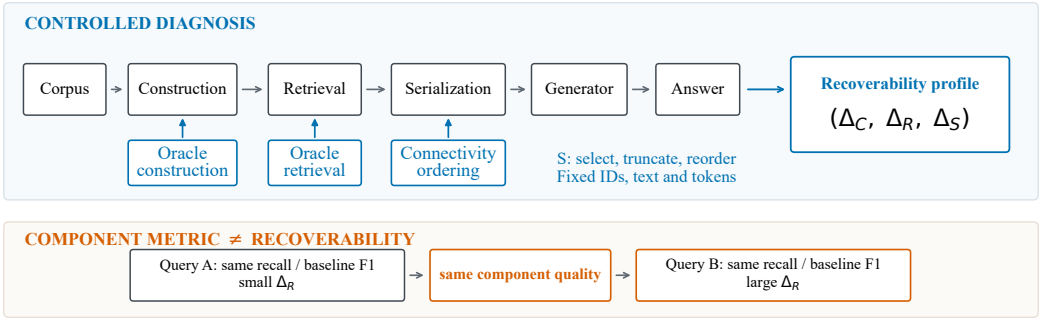

Figure 1: INTERVENERAG measures downstream responses to specified construction, retrieval, and ordering interventions. The pure S probe permutes only the final retained evidence, after selection and truncation, preserving IDs, text, and token count. The lower illustration contrasts component quality with repair response.

ordering probes test presentation sensitivity, with downstream operators and the generator fixed. Our primary retrieval intervention selects existing graph evidence within automatic unit/token budgets. Joint interventions test interactions, and frozen-list prefixes compare smaller annotation-guided selections with larger automatic selections.

Across two datasets and three generator families, in-graph selection improves F1 by 0.206–0.339 despite high annotated-fact coverage. Five component statistics weakly predict these responses under the tested linear and nonlinear models. On 2Wiki, graph-size and selected-context controls attenuate historical construction–retrieval interactions by 67–81%; smaller annotation-guided prefixes often outperform the largest automatic prefix tested.

The primary study uses per-query benchmark candidates. Construction repair is defined only on 2Wiki, whose relational annotations map to the graph schema; retrieval and ordering cover both datasets. Extensions test MuSiQue, shared HippoRAG indices, and a fully bounded factorial. Gains remain conditional on the replacement, without guaranteeing deployable improvement.

Our contributions are:

- A stage-level framework measuring paired answer responses to construction, retrieval, and ordering interventions.
- Controlled comparisons exposing candidate-space and evidence-budget constraints and their effects on gains and interactions.
- Cross-generator evidence of in-graph selection headroom and protocol-sensitive diagnosis, with additional evidence-interface and architecture checks.

2 RELATED WORK

2.1 GRAPH-BASED RETRIEVAL-AUGMENTED GENERATION

Graph-based RAG connects evidence across sources through entity and passage relations. HippoRAG uses a knowledge graph and personalized PageRank for associative retrieval (Jiménez Gutiérrez et al., 2024), and HippoRAG 2 integrates passage information into this memory (Jiménez Gutiérrez et al., 2025). LightRAG combines graph and vector retrieval at entity and broader relational levels (Guo et al., 2024). Community-summary GraphRAG targets global query-focused summarization (Edge et al., 2024); Think-on-Graph iteratively explores a knowledge graph (Sun et al., 2024); G-Retriever studies retrieval for textual graphs (He et al., 2024). Our stage decomposition instantiates one such modular pipeline, and INTERVENERAG measures final-answer responses to specified construction, selection, and ordering interventions within it.

2.2 Evaluation and Intervention-Based Diagnosis

RAGAs and ARES assess context relevance, answer faithfulness, and answer relevance (Es et al., 2024; Saad-Falcon et al., 2024); RAGChecker diagnoses retrieval and generation through claim-level entailment (Ru et al., 2024). GraphRAG evaluations additionally assess construction and reasoning across tasks and domains (Xiang et al., 2025; Xiao et al., 2025).

Evidence interventions measure answer responses to changed context. Xia (2026) compares no-evidence, retrieved, full-context, and oracle conditions; Bio-GRACE measures evidence benefit in biomedical verification (Deka & Singh, 2026); QO-Bench separates retrieval from query-operator failures using typed-event questions and oracle evidence (Zhang et al., 2026); ARC-JSD ablates sentences for response-distribution attribution (Li et al., 2026).

InterveneRAG examines GraphRAG stages individually and jointly, controlling graph size and context budgets to relate paired answer gains to each replacement's candidate evidence and exposure. Annotations provide diagnostic references, not guaranteed ceilings; Appendix F distinguishes this target from general attribution.

2.3 Evidence Selection and Context Budgets

Retrieval policies adapt evidence acquisition and selection. Self-RAG learns on-demand retrieval and critiques passages and responses (Asai et al., 2024). Adaptive-RAG chooses no retrieval, single-step retrieval, or iterative retrieval by question complexity (Jeong et al., 2024). LDAR balances useful evidence, distracting context, and token usage through learned retrieval (Shim et al., 2026). These methods optimize inference-time policies. Our budget experiments complement them by quantifying how much selection headroom remains inside an existing graph: with graph and search fixed, automatic and annotation-guided list prefixes compare better selection against more automatic evidence.

3 InterveneRAG: Controlled Stage Interventions

InterveneRAG measures answer-quality changes after intervening on evidence construction, selection, or ordering. An intervention is the stage operation; a replacement is its alternative output; a repair replaces one stage's output; and its gain is the answer-quality change relative to the automatic baseline.

3.1 Pipeline and Diagnostic Setup

For query x and its candidate documents, construction C produces a provenance-linked entity–relation graph, retrieval R selects evidence units, and serialization S forms an ordered context for a fixed generator. Diagnosis obtains an automatic baseline, replaces a specified stage output, and reruns downstream operators with unchanged settings (Figure 1). Construction and retrieval replacements use supporting-fact annotations, available only to the diagnostic replacements; ordering uses graph connectivity. The comparison measures answer responses without requiring a learned replacement policy.

3.2 Stage-Specific Interventions

Construction. We replace the automatic graph with one derived from annotated supporting facts, retaining the same graph schema, retrieval operator, serializer, and generator. This intervention tests how answer quality changes when downstream processing receives an annotation-derived graph. It is defined for 2WikiMultiHopQA, whose annotations support a mapping into the existing representation.

In-graph retrieval. We freeze the automatic graph $G_{\text{auto}}(x)$ and use supporting-fact annotations to prioritize evidence already present in it. The selected evidence identities satisfy

$$R^*_{\text{in-graph}}(x) \subseteq U(G_{\text{auto}}(x)), \qquad K^*(x) \le K_{\text{auto}}(x), \qquad B^*(x) \le B_{\text{auto}}(x), \tag{1}$$

where $U(G)$ is the set of graph evidence units, K counts evidence units and B counts canonical tokens in the serialized context. The selector prioritizes distinct supporting facts and fills remaining

capacity with automatically ranked evidence; source text is truncated if needed to satisfy the token ceiling. The serializer and generator remain fixed. This intervention tests the gains available from selecting evidence within the existing graph without increasing either the unit count or token budget.

Evidence ordering. We retain the selected evidence units and their contents, replacing only their order with a deterministic connectivity-based traversal. The generator therefore receives the same evidence in a different sequence. This intervention tests sensitivity to presentation; the traversal is a probe rather than a tuned ordering.

Selection rules, annotation mappings, and intervention invariants are detailed in Appendix B.2.

3.3 Measuring Recoverability

For the same query x, let $m_0(x)$ be the automatic answer score and $m_j^{\text{int}}(x)$ the score after intervening on stage $j \in \{C, R, S\}$. Stage-level recoverability is the paired difference

$$\Delta_j(x) = m_j^{\text{int}}(x) - m_0(x). \tag{2}$$

For example, an increase in answer F1 from 0.30 to 0.70 gives $\Delta_R(x) = 0.40$. Positive values indicate improvement under the specified intervention, zero indicates no change, and negative values indicate degradation. A zero gain means this replacement did not help; it does not show that the stage needs no repair.

The three gains form a query's recoverability profile. Each gain is conditional on the replacement and the fixed comparison settings. It measures an observed diagnostic response, rather than intrinsic component responsibility or the expected benefit of a deployable policy.

3.4 Joint Interventions and Comparison Controls

Two individually useful interventions need not yield additive improvements. Construction and retrieval may recover overlapping answer quality, while replacing the graph also changes the evidence that retrieval can access. To examine this dependence, we evaluate all eight combinations of automatic and intervened construction, retrieval, and ordering on 2WikiMultiHopQA.

The historical factorial uses a distinct retrieval protocol, denoted R^{uncapped} rather than the budget-constrained $R^*_{\text{in-graph}}$. At construction level c, automatic retrieval runs the default search on G_c; intervened retrieval retains all units linked to annotated supporting facts in that graph, without the single-stage selector's count/token ceilings. For $c, r, s \in \{0, 1\}$, with 0 denoting automatic operation and 1 the specified intervention, the answer score is

$$Q_x(c, r, s) = m(\text{Gen}(S_s(R_r(G_c(x))))), \tag{3}$$

where m scores the generated answer against the reference and the query input to generation is implicit. Holding ordering at its automatic setting, we measure

$$I_{CR}(x) = Q_x(1, 1, 0) - Q_x(1, 0, 0) - Q_x(0, 1, 0) + Q_x(0, 0, 0). \tag{4}$$

A negative value means that the joint gain is smaller than the sum of the separate gains. Its magnitude also depends on the candidate graph and evidence exposure. We therefore repeat the comparison with repaired graphs matched to automatic edge counts and, in a stricter control, with context exposure constrained by the automatic token budget. The historical factorial appears in Appendix E. A fresh eight-cell replication instead caps every cell at the automatic unit/token budget, prioritizes distinct supporting facts for $R = 1$, and applies S only after truncation (Appendix H). These protocols define different estimands.

Beyond stage combinations, we examine changes in evidence budget. To index these comparisons, let A denote the intervened stage set, e the budget or operating point, and ρ the intervention protocol. Define

$$Q_x(A, e; \rho) = m(\text{Repair}(x, A, e; \rho)), \qquad V_x(A, e; \rho) = Q_x(A, e; \rho) - Q_x(\varnothing, e), \tag{5}$$

where Repair returns the generated answer and $Q_x(\varnothing, e)$ is the corresponding automatic baseline. Together, these outcomes form a recoverability surface over intervention combinations and comparison conditions. Section 3.5 examines its budget dimension with a fixed graph and search configuration. Table 1 indexes the protocols used in the results.

Table 1: Protocols referenced in the main text. Every protocol fixes the extraction prompt, serializer, prompt template, and generator; they differ in how replacements are built and in which ceilings and matching conditions apply. "Snapshot" marks whether a protocol belongs to the historical run set or to the fresh extensions; values from different snapshots are not pooled.

Protocol	Construction C	Retrieval R and ordering S	Ceiling and matching	Snapshot
P1	automatic graph, frozen	annotation-prioritized selection inside G_{auto}; automatic ranking fills the remaining capacity	at most the automatic unit and token counts	historical
P2	automatic or annotation-derived (G_{oracle})	default search, or every unit linked to annotated supporting facts; ordering automatic	none	historical
P3	as P2, with repaired graphs matched to automatic edge counts	as P2	300 queries; edges matched, composition and exposure not matched	historical
P4	as P3	as P2, but the selected C_1R_0 context is capped	automatic token budget; 8 token-infeasible queries excluded	historical
P5	automatic or annotation-derived	default search, or distinct-fact priority; ordering applied after truncation	every cell capped at the automatic unit and token budget	fresh
P6	automatic graph, frozen	prefix of the frozen automatic list or of the annotation-guided list	search fixed at R16; final unit cap $k \in \{4, 8, 12, 16\}$	historical and fresh

3.5 INTERVENERAG: BUDGETED REPAIR UNDER FIXED SEARCH

We compare more automatic evidence with a smaller annotation-guided selection, freezing the graph and automatic baseline search. The annotation-guided selector accesses the whole graph, including evidence outside the automatic search output.

For each query, automatic retrieval runs once with the broader R16 configuration. We freeze its ordered output and an annotation-guided list from the same graph, then serialize their first k units (or all remaining units) for $k \in \{4, 8, 12, 16\}$. This final evidence-unit cap does not change search depth. Within each list, construction, search, scoring, order, serializer, prompt, and generator remain fixed.

Same-cap comparisons measure composition and ordering effects under a common unit limit. The primary comparison gives automatic selection the larger cap: annotation-guided $k \in \{4, 8, 12\}$ versus automatic $k = 16$. In Equation 5's notation, this is

$$Q_x(\{R\}, (\text{R16}, k); \rho_{\text{in}}) - Q_x(\varnothing, (\text{R16}, 16)), \tag{6}$$

where ρ_{in} denotes the fixed-search, annotation-guided in-graph selection protocol.

A positive difference identifies selection headroom under a smaller unit cap. Units vary in length, so a smaller unit cap does not imply fewer tokens, and these comparisons do not bound total system cost. Appendix Table 19 reports realized context-token counts as evidence exposure.

4 EXPERIMENTS

We examine four questions: whether component metrics predict recoverability, how much selection headroom remains in existing graphs, how construction and retrieval interact, and whether increasing evidence caps closes the selection gap.

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate on 2WikiMultiHopQA and HotpotQA, multi-hop QA benchmarks with supporting-fact annotations. We sample 300 validation queries per dataset, stratified by question type, and reuse these cohorts across comparisons. Graphs are built from each query's benchmark-provided candidate context, using HotpotQA's distractor setting. The primary evaluation concerns per-query

candidate graphs. Extensions add 300 MuSiQue questions (100 each with two, three, and four hops) and strictly indexed selection on shared HippoRAG corpora; their supervision and candidate scopes are defined in Appendix H.

Pipeline and generators. The primary pipeline extracts subject–relation–object triples with DeepSeek V4 Flash, retaining source-sentence provenance. Retrieval combines dense similarity, BM25, and graph connectivity, and selected evidence is serialized with source text. DeepSeek V4 Flash, Qwen3-8B, and GLM-5.2 generate answers from this shared pipeline. Extraction prompts, retrieval parameters, and model configurations appear in the appendix.

Comparison protocol. Contrasts pair queries within each generator against the automatic baseline using Section 3's interventions. Prompt templates, generation settings, answer normalization, and downstream operators otherwise remain fixed. Main stage repairs cover all three generators. Construction–retrieval confirmations, matched controls, and fixed-search comparisons use DeepSeek and GLM with three fresh generations per condition. Qwen stage comparisons use one. A historical Qwen HippoRAG 2 interface check is supplemented by fresh two-generator indexed-passage comparisons. The new eight-cell bounded factorial and order controls use the predefined 292-query 2Wiki cohort; every cell has three fresh generations. Table 18 summarizes the experiments.

Metrics and uncertainty. We report paired changes in token-overlap F1 and exact match, with 95% bootstrap intervals over queries. Repeated-generation comparisons use hierarchical resampling within queries. Significance tests treat queries as independent units, with Holm correction within each prespecified fixed-search comparison family and outcome. Appendices D.2 and H specify resampling and comparison families for historical and fresh experiments, respectively.

4.2 Component Metrics and Recoverability

We test whether five automatic-pipeline measurements predict query-level in-graph F1 gains: support recall, graph support coverage, evidence-unit count, token count, and annotation-nonsupport count. Annotation-derived features are offline diagnostic signals. All three predictors in Figure 2A use identical five outer query folds, with hyperparameters selected in three inner folds.

Across six dataset–generator settings, nested-CV R^2 is 0.091–0.138 for ridge, 0.099–0.146 for quadratic ridge, and 0.028–0.109 for boosting. Thus, these five statistics and the three predictor classes tested leave most query-level variation unexplained. The practical implication is that in-graph headroom is tied to which evidence is selected rather than to aggregate component statistics; richer features or other predictor classes may explain part of the remainder. The historical linear analysis, baseline-F1 augmentation, and selected metric-matched case illustrations remain in Appendix G and Figure 6.

4.3 In-Graph Selection Headroom

Across six dataset–generator settings, Δ_R (Equation 2) compares automatic retrieval with the annotation-guided selector of Equation 1 (P1, Table 1), fixing the graph and generator without increasing evidence-unit or canonical token counts. Mean gains are 0.206–0.339, with all paired 95% bootstrap intervals above zero (Figure 2, right). Automatic graphs already cover 94.9% of annotated supporting facts on 2Wiki and 96.6% on HotpotQA within benchmark candidate pools. High availability therefore leaves substantial selection headroom within existing graphs and budgets.

Two checks examine selector and architecture sensitivity. Removing source duplicates changes gains by -0.026 to $+0.004$ across the six settings (Table 6), preserving the headroom. With Qwen3-8B on HippoRAG 2, token-matched interface gains are 0.136 on 2Wiki and 0.080 on HotpotQA (Table 9), extending the evidence beyond the primary triple-extraction architecture. This historical interface comparison is distinct from the fresh strict indexed-passage experiment reported below.

4.4 Construction–Retrieval Interactions

We next test whether construction and retrieval gains add up. On 2WikiMultiHopQA, we compare the four construction–retrieval combinations with automatic ordering, using DeepSeek and GLM with

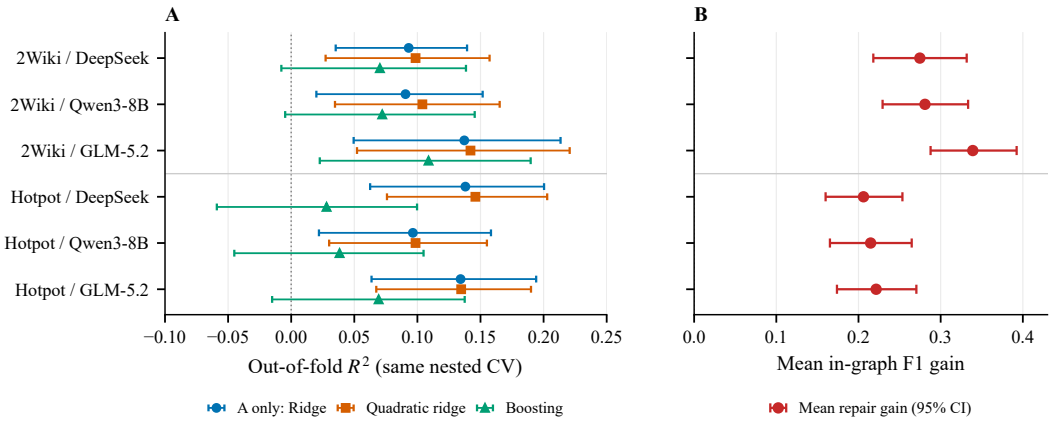

Figure 2: A: colors identify component-only predictors on identical nested query folds. The 95% intervals bootstrap fixed out-of-fold predictions and exclude retraining uncertainty. B: red points show historical mean in-graph repair gains over automatic retrieval; bars are paired 95% intervals. Graph identities and automatic unit/token ceilings are fixed. Panels measure distinct quantities.

three paired generation replicates. Following Section 3, retrieval intervention retains all annotated-support evidence in each construction condition's graph; this uncapped factorial (P2, Table 1) differs from the budget-constrained selector above.

Interactions of this form are bounded above: $I_{CR} \leq 1 - Q(1,0,0) - Q(0,1,0) + Q(0,0,0)$. Under the uncapped protocol that bound is already -0.280 for DeepSeek, so a negative value is not by itself evidence that the two stages substitute for one another. The observed interactions are -0.321 (DeepSeek) and -0.340 (GLM); absolute cell means appear in Appendix Table 23.

Two controls show how much of that negativity survives matched comparison conditions. Matching repaired graphs to automatic edge counts on 300 queries (P3) yields interactions of -0.086/-0.111 (DeepSeek/GLM). The stricter control (P4) additionally caps only the selected C_1R_0 context while preserving its annotation-derived core, and leaves the other cells intact. On the 292 feasible queries, excluding eight whose core exceeds the automatic token ceiling, interactions are -0.067/-0.066, against -0.319/-0.343 under the same-cohort uncapped protocol. Edge matching also changes composition and exposure, so we read the attenuation as protocol-dependent and quantify the fully bounded case next.

The fresh eight-cell protocol (P5) caps every cell at the automatic unit and token budget, prioritizes distinct supporting facts for $R = 1$, and applies ordering after truncation. It puts the interaction at -0.024/$+0.002$, with both intervals crossing zero (Figure 3), and none of the six pure-order F1 contrasts is significant after Holm correction. Because P5 changes the retrieval rule, the model snapshot, and the evidence ceiling together, it estimates a different quantity from P2–P4: it asks whether construction–retrieval interaction persists when every cell faces the same ceiling, and it is not by itself a test of stage independence. All cells and higher-order terms appear in Appendix H.

4.5 Evidence Selection under Varying Caps

We finally test whether more automatic evidence closes the gap to annotation-guided selection. Following Section 3.5 (P6, Table 1), we fix the graph and the R16 search and take prefixes of the frozen automatic and annotation-guided lists at caps $k \in \{4, 8, 12, 16\}$. The annotation-guided list can use the complete graph. Both datasets use DeepSeek and GLM, with 300 paired queries and three generation replicates per setting.

Automatic selection generally improves with the cap, but the gap remains (Figure 4). Annotation-guided curves rise from $k = 4$ to 8 and largely flatten on 2Wiki, while remaining relatively stable on HotpotQA. Same-cap F1 gains span 0.130–0.257 across 16 dataset–generator–cap settings. All 16 comparisons significantly favor annotation-guided selection for F1 and exact match after per-outcome

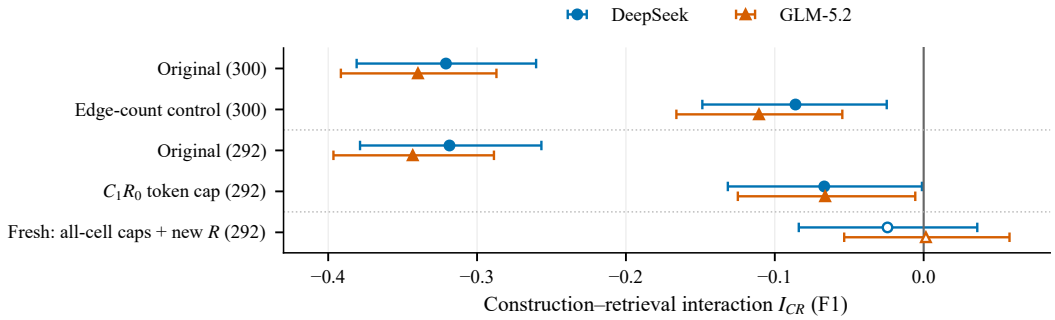

Figure 3: Construction–retrieval interactions with unified mean-of-three 95% intervals; parentheses give query counts. Separators distinguish cohorts and the fresh protocol. The historical token control caps only C_1R_0; the fresh protocol caps all cells and changes R to distinct-fact priority. These are distinct comparison protocols, not a single-factor ablation chain.

Holm correction (Appendix Table 15), measuring composition and ordering effects under a common unit limit.

Giving automatic selection the larger cap, we compare annotation-guided prefixes at $k \in \{4, 8, 12\}$ against automatic $k = 16$. Annotation-guided selection is better in 11 of 12 comparisons after per-outcome Holm correction (Appendix Table 16). The single exception is 2Wiki–DeepSeek at $k = 4$, where the F1 gain is $+0.039$ (95% CI $[-0.015, +0.093]$) and the exact-match gain is $+0.034$ ($[-0.020, +0.090]$). That prefix covers all annotated facts in only 61% of 2Wiki queries and repeats source facts, which points to a coverage and ordering limit of the $k = 4$ repair rather than to a need for more than four distinct facts. At $k = 8$ and $k = 12$, annotation-guided selection wins consistently.

Within the frozen graph, raising the automatic cap to 16 does not consistently close the gap to annotation-guided selection: choosing evidence with annotations yields better answers under a smaller unit cap. The observed Δ_R values quantify how much answer quality remains recoverable under the tested selection and budget conditions and, as noted in Section 3.5, they are not token or cost bounds. The natural next step is a selector that reaches comparable gains without access to supporting-fact annotations; these numbers define the target it has to hit.

In a fresh equal-cap $k = 4$ comparison, distinct-fact-first selection from the same repaired list raises complete support coverage from 0.610 to 0.827 and adds $+0.129$/$+0.138$ F1 (DeepSeek/GLM). This changes the selected content rather than its presentation; token counts are not matched (Appendix H).

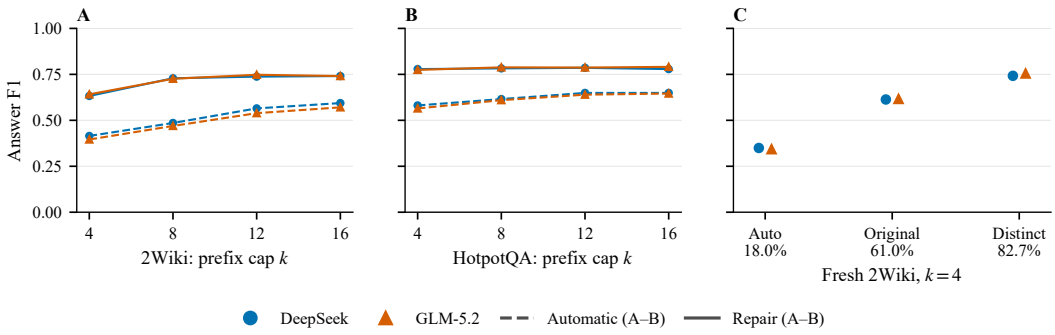

Figure 4: A–B: historical three-generation F1 curves over frozen source-list prefixes. C: separate fresh 2Wiki $k = 4$ means for automatic, original-repair, and distinct-fact-first selection; percentages below labels show complete annotated-support coverage. This changes selected content and does not strictly match tokens. New calls are not pooled with historical curves; paired contrast intervals appear in Appendix H.

4.6 Extensions across Tasks and Shared Indices

The four questions above concern the primary cohorts. We next ask whether the same headroom appears under a different dataset and a shared index. On the equal-hop MuSiQue cohort, paragraph-guided graph selection changes F1 by $+0.293$ for DeepSeek and $+0.305$ for GLM. Supervision here is paragraph-level, so a paragraph that appears in the annotation need not contain every necessary relation, and these gains should be read as selection headroom rather than as an extraction guarantee. No-context and missing-support controls, hop strata, and absolute scores appear in Appendix H.

Strictly indexed HippoRAG passage selection changes F1 by $+0.086$–$+0.117$ across the two datasets and generators (Table 2). Every selected passage is verified against the frozen graph and embedding store, and repair stays within the native automatic passage/token ceilings. The shared corpora pool benchmark candidates and contain the mapped support; this extends the evidence to bounded shared indices without establishing open-corpus candidate acquisition.

Table 2: Fresh extension experiments, 300 queries per dataset and three generations per condition. Repair selects existing graph units or indexed passages under the corresponding automatic unit/token ceilings. Brackets are marginal paired 95% intervals; these are separate model snapshots from historical comparisons.

Setting	Dataset	Generator	Auto F1	Repair F1	Gain [95% CI]
Paragraph graph	MuSiQue	DeepSeek	0.306	0.599	$+0.293$ [$+0.243$, $+0.345$]
Paragraph graph	MuSiQue	GLM-5.2	0.302	0.607	$+0.305$ [$+0.252$, $+0.358$]
Shared HippoRAG	2Wiki	DeepSeek	0.652	0.768	$+0.117$ [$+0.080$, $+0.154$]
Shared HippoRAG	2Wiki	GLM-5.2	0.698	0.783	$+0.086$ [$+0.056$, $+0.117$]
Shared HippoRAG	HotpotQA	DeepSeek	0.683	0.791	$+0.109$ [$+0.077$, $+0.143$]
Shared HippoRAG	HotpotQA	GLM-5.2	0.703	0.795	$+0.092$ [$+0.059$, $+0.126$]

5 Discussion and Limitations

GraphRAG evaluation should pair component metrics with controlled answer responses. Graphs covering about 95% of annotated facts still permit substantial in-graph selection gains, while construction–retrieval interactions vary from about -0.34 to near zero across graph and exposure controls. We therefore recommend reporting recoverability alongside the replacement rule, candidate source, and evidence ceiling. These gains are replacement-specific empirical references, not deployable ceilings or causal stage attributions. Primary cohorts use per-query benchmark candidates; MuSiQue adds paragraph-level supervision, and HippoRAG uses shared indices of benchmark candidates. Missing-document recovery, full-corpus candidate acquisition, and construction replacements without an architecture-preserving mapping remain outside scope. Architecture and corpus scope vary together in the HippoRAG extension. Construction and factorial evidence is limited to 2Wiki; the fresh bounded factorial uses its predefined 292-query feasible cohort. Provenance and paragraph coverage cannot certify semantic completeness, especially after truncation. The prediction study covers five measurements and linear, quadratic, and boosting models; richer features and other predictors remain untested. Budget comparisons stop at $k = 16$. Although new runs record provider token usage and call timing, efficiency claims require matched end-to-end cost studies including construction and indexing.

6 Conclusion

We introduced InterveneRAG to measure GraphRAG recoverability through paired answer-quality changes under controlled stage interventions. Across two multi-hop QA datasets and three generators, annotation-guided in-graph selection improves mean F1 by 0.206–0.339 within automatic evidence-unit and token budgets. Construction–retrieval interactions depend on graph and context matching, while smaller annotation-guided selections often outperform the largest tested automatic cap under fixed search. Fresh MuSiQue, shared-index HippoRAG, and fully bounded factorial experiments broaden the tested conditions while preserving their supervision and candidate-scope boundaries. These findings support intervention-based diagnosis with explicit comparison conditions.

AI USE STATEMENT

Generative AI tools were used to assist with literature search, manuscript drafting, language editing, experimental code development, and experimental execution. AI agents assisted in implementing experimental code and carrying out experimental workflows under the authors' supervision. The authors reviewed the AI-assisted work, verified the reported experimental results, and take full responsibility for the research and final manuscript.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Pritam Deka and Prabhjot Singh. When retrieval helps and distracts: Evaluating evidence-generating LLMs for biomedical claim verification, 2026. URL https://arxiv.org/abs/2608.01409.

Darren Edge, Ha Trinh, Newman Cheng, et al. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158, 2024. doi: 10.18653/v1/2024.eacl-demo.16. URL https://aclanthology.org/2024.eacl-demo.16/.

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

Xiaoxin He, Yijun Tian, Yifei Sun, et al. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. In *Advances in Neural Information Processing Systems*, 2024.

Kosuke Imai, Luke Keele, and Dustin Tingley. A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334, 2010. doi: 10.1037/a0020761.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity, 2024. URL https://arxiv.org/abs/2403.14403.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=hkujvAPVsg.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models, 2025. URL https://arxiv.org/abs/2502.14802.

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*, 2020.

Ruizhe Li, Chen Chen, Yuchen Hu, Yanjun Gao, Xi Wang, and Emine Yilmaz. Attributing response to context: A Jensen–Shannon divergence driven mechanistic study of context attribution in retrieval-augmented generation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2505.16415.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

Judea Pearl. Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pp. 411–420, 2001.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems*, 2024.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 338–354, 2024. doi: 10.18653/v1/2024.naacl-long.20. URL https://aclanthology.org/2024.naacl-long.20/.

Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games II*, pp. 307–317. Princeton University Press, 1953.

Seongwoong Shim, Myunsoo Kim, Jae Hyeon Cho, and Byung-Jun Lee. Beyond RAG vs. long-context: Learning distraction-aware retrieval for efficient knowledge grounding. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.21865.

Ilya M. Sobol'. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiment*, 1:407–414, 1993.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, et al. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *International Conference on Learning Representations*, 2024.

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/.

Haizhou Xia. Diagnosing evidence utilization in long-context and retrieval-augmented language models under matched evidence conditions, 2026. URL https://arxiv.org/abs/2606.06758.

Zhishang Xiang, Chuanjie Wu, Qinggang Zhang, Shengyuan Chen, Zijin Hong, Xiao Huang, and Jinsong Su. When to use graphs in RAG: A comprehensive analysis for graph retrieval-augmented generation, 2025. URL https://arxiv.org/abs/2506.05690.

Yilin Xiao, Junnan Dong, Chuang Zhou, Su Dong, Qian-wen Zhang, Di Yin, Xing Sun, and Xiao Huang. GraphRAG-Bench: Challenging domain-specific reasoning for evaluating graph retrieval-augmented generation, 2025. URL https://arxiv.org/abs/2506.02404.

Mengao Zhang, Xiang Yang, Chang Liu, Tianhui Tan, and Ke-wei Huang. QO-Bench: Diagnosing query-operator-preserving retrieval over typed event tuples, 2026. URL https://arxiv.org/abs/2606.04646.

Table 3: InterveneRAG repair coordinates. Each row identifies a controlled repair or analysis axis, its invariants, and the scope of the resulting recoverability measurement.

Coordinate	What changes	Held fixed	Interpretation boundary
Construction C	Graph evidence source G_c	Schema, operator definitions, serializer, generator	2Wiki only; opportunity set may change
In-graph retrieval R	Selection within G_{auto}	Automatic graph, count/token ceilings, serializer, generator	Primary stage-level measurement
Source-deduplicated R	Deduplicated support-first selection	Automatic graph, count/token ceilings, generator	Distinct graph-internal variant
Historical factorial R	Provenance filter on current G_c	Operator rule, serializer, generator	Candidate graph depends on C; repaired output is uncapped
Fixed-search selection	Annotation-guided ordering within G_{auto}	Automatic graph, R16 source lists, serializer, generator	Exposure analysis; uses gold relevance
Serialization S	Deterministic evidence order	Evidence identity, generator, prompt	Position sensitivity only
Operating point o	Seeds, hops, maximum edges	Scoring weights, graph, prompt, generator	R4/R12/R16 sensitivity; factors change jointly
Prefix cap k	Final serialized units	Graph, R16 search and source-list order, prompt, generator	Online exposure only; not end-to-end cost

A Reproducibility Protocol

All analyses operate on fixed artifacts. Each controlled comparison is paired by query and generator, changes only the named intervention field, and evaluates the same downstream answer metric. Oracle conditions provide empirical diagnostic references under the specified protocol and are not guaranteed performance upper bounds.

Data and splits. The 2Wiki cohort is drawn from 12,576 validated official-dev examples. The HotpotQA cohort is drawn from the distractor validation split after one row with an unresolved supporting-fact sentence index is excluded, leaving 7,404 eligible examples. Both 300-query cohorts use the same deterministic, question-type-stratified sampler with seed 42 rather than selecting the first 300 rows; the HotpotQA cohort contains 240 bridge and 60 comparison questions. Ordered query IDs are recorded in the manifests.

Primary GraphRAG pipeline. Each source paragraph is processed by `deepseek-v4-flash` with thinking disabled, temperature 0, JSON-object mode, and the recorded `compact_v2` prompt. Each extracted subject–relation–object triple stores an exact source title, zero-based sentence index, source text, and confidence. We retain an edge only when its index is in range, its source text matches the sentence after whitespace normalization, and its relation fields are nonempty.

In the original factorial, automatic retrieval ranks all N extracted edges by

$$s_i = .55(1 - r_i^{\text{dense}}/N) + .25(1 - r_i^{\text{BM25}}/N). \tag{7}$$

BM25 uses $k_1 = 1.5$ and $b = .75$; dense retrieval uses normalized `BAAI/bge-small-en-v1.5` embeddings with dot-product scoring. The top two edges seed up to two connectivity hops with a 0.20 bonus, and at most eight edges are retained. We denote this default operator by R8=$(2, 2, 8)$. The factorial repair instead returns every gold-provenance edge in the current construction graph and imposes no eight-edge cap. Ties are resolved by original edge index. The serializer emits each edge followed by its exact source. The ordering intervention keeps the evidence multiset and all representation fields fixed, changing only a deterministic graph-connectivity traversal. The 2Wiki construction oracle maps native annotated triples into the same representation; both R8 search and the uncapped provenance filter operate on the graph supplied by the construction level. HotpotQA has no corresponding architecture-preserving construction condition.

The exposure study separately defines R16=$(4, 3, 16)$ on the frozen automatic graph. It constructs the automatic-search and annotation-guided in-graph source lists once, then applies final serialization caps $k \in \{4, 8, 12, 16\}$. R16 is not the factorial retriever, and k is not a maximum-edge parameter.

Provenance. Oracle evidence is mapped to annotated source material before retrieval or generation. Fixed manifests record source artifacts, condition identifiers, query counts, and hashes. The evidence ledger provides a machine-readable source key for every quantitative claim.

B Controlled Retrieval and Context Analyses

B.1 Fixed-Automatic-Graph Selection Protocol

For the exposure and constrained-repair studies, the frozen automatic graph is the complete candidate universe for each query. Gold supporting facts are mapped only through exact title/index provenance after source-text validation. The oracle selects one unit for each represented supporting fact, breaks

ties by frozen automatic rank and stable ID, and fills the remaining budget with support-relevant and automatic-ranked units under the stated unit-count and canonical-token ceilings. Every selected unit is checked against the frozen graph; no graph construction, OpenIE call, new scorer, or graph-external support is introduced. This protocol is distinct from the factorial provenance filter applied to condition-specific G_c.

Mean exact support coverage is 0.949 on 2Wiki and 0.966 on HotpotQA, with 267 and 277 coverage-complete queries, respectively. Coverage correlations are descriptive.

B.2 Source-Deduplicated Constrained In-Graph Variant

For automatic items A and the full gold-provenance filter G drawn from the same automatic graph, let $K_{\text{auto}} = |A|$ and let B_{auto} denote the canonical-token count of the serialized automatic context. The deterministic repair proceeds as follows:

1. returns an empty context when $K_{\text{auto}} = 0$; otherwise traverses G followed by A, deduplicating by title, sentence index, and source text;
2. retains the first K_{auto} candidates, prioritizing support and using nonduplicate automatic-ranked units as filler;
3. applies the same serializer and, when necessary, truncates the final source text at a token boundary so that $B_{\text{repair}} \leq B_{\text{auto}}$; and
4. verifies both ceilings, hashes the context, and uses the original prompt and generator without rerunning construction or retrieval.

This is the source-deduplicated variant, distinct from the primary selector that prioritizes one edge per represented supporting fact before adding ranked fillers. All 600 stored variant contexts use existing automatic-graph IDs. Both variants use the same annotated-fact recall denominator in Table 4; historical oracle-relative retention values are not directly comparable to total graph coverage. Unit-count ceilings need not be filled: deduplication and token truncation reduce the selected count. Residual differences include semantic density, order, and context composition.

Table 4: Context composition using one recall definition: represented distinct annotated (title, sentence-index) facts divided by all annotated supporting facts. Values are means (medians). Both constrained variants select existing automatic-graph IDs. Provenance recall does not certify semantic completeness after text truncation.

Dataset	Condition	Units	Canonical tokens	Support recall	Nonsupport
2Wiki	Automatic	6.70 (8.0)	347.7 (371.0)	0.558	4.12
2Wiki	Full in-graph provenance filter	4.64 (4.0)	249.7 (188.5)	0.949	0.00
2Wiki	Source-deduplicated selection	5.20 (5.0)	266.8 (262.5)	0.945	2.88
2Wiki	Support-first in-graph selection	6.47 (7.0)	335.5 (358.0)	0.945	2.54
HotpotQA	Automatic	6.64 (8.0)	363.8 (388.0)	0.604	4.30
HotpotQA	Full in-graph provenance filter	3.98 (3.0)	228.0 (186.0)	0.966	0.00
HotpotQA	Source-deduplicated selection	5.24 (5.0)	283.3 (285.5)	0.959	2.90
HotpotQA	Support-first in-graph selection	6.32 (7.0)	350.1 (376.5)	0.957	2.70

Table 5: Source-deduplicated count- and token-constrained in-graph selection on 300 queries per dataset. The reported fraction divides the constrained gain by the full-oracle gain from the same graph-serializer lineage and is descriptive.

Dataset	Generator	Auto F1	Constrained F1	Constrained gain (95% CI)	Full-oracle gain	Fraction
2Wiki	DeepSeek	0.465	0.744	+0.279 [+0.222, +0.336]	+0.302	0.922
HotpotQA	DeepSeek	0.589	0.788	+0.199 [+0.155, +0.243]	+0.230	0.863
2Wiki	Qwen3-8B	0.387	0.642	+0.255 [+0.202, +0.307]	+0.293	0.869
HotpotQA	Qwen3-8B	0.515	0.708	+0.193 [+0.145, +0.242]	+0.236	0.818
2Wiki	GLM-5.2	0.422	0.765	+0.343 [+0.291, +0.397]	+0.349	0.983
HotpotQA	GLM-5.2	0.560	0.776	+0.216 [+0.170, +0.264]	+0.242	0.894

Table 6: Controlled effects in the primary GraphRAG pipeline. The primary support-first and source-deduplicated variants both select automatic-graph units, but use different ordering, deduplication and filler rules. Entries are paired answer-F1 gains with 95% bootstrap intervals.

Dataset	Generator	Construction	In-graph retrieval	Source-deduplicated	Serialization
2Wiki	DeepSeek	+0.487 [+0.429, +0.545]	+0.275 [+0.218, +0.332]	+0.279 [+0.222, +0.336]	-0.013 [-0.033, +0.003]
2Wiki	Qwen3-8B	+0.482 [+0.425, +0.538]	+0.281 [+0.229, +0.333]	+0.255 [+0.202, +0.307]	-0.004 [-0.028, +0.019]
2Wiki	GLM-5.2	+0.470 [+0.414, +0.526]	+0.339 [+0.288, +0.392]	+0.343 [+0.291, +0.397]	-0.007 [-0.020, +0.007]
HotpotQA	DeepSeek	—	+0.206 [+0.160, +0.253]	+0.199 [+0.155, +0.243]	+0.001 [-0.012, +0.014]
HotpotQA	Qwen3-8B	—	+0.215 [+0.165, +0.265]	+0.193 [+0.145, +0.242]	-0.027 [-0.048, -0.008]
HotpotQA	GLM-5.2	—	+0.221 [+0.174, +0.270]	+0.216 [+0.170, +0.264]	+0.001 [-0.018, +0.019]

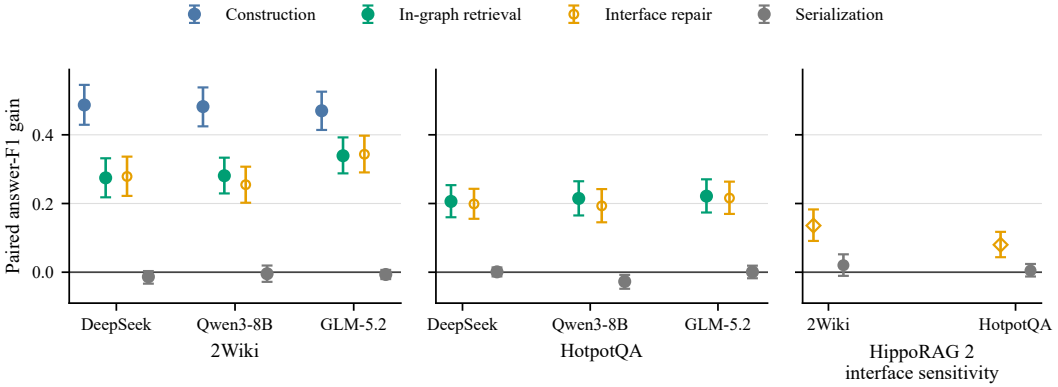

Figure 5: Better evidence selection within the existing graph improves answer F1. Left and center show the primary in-graph repair; right shows the token-matched HippoRAG 2 interface comparison. Filled green markers denote in-graph conditions, hollow orange markers denote interface conditions, and error bars are 95% paired bootstrap intervals.

B.3 Complete Stage-Level Repair Effects

The complete six-setting stage-level comparison reports construction, support-first in-graph retrieval, source-deduplicated selection, and serialization repairs under their respective held-fixed conditions.

B.4 Metric-Matched Discordant Cases

The matched-pair analysis supplements predictive R^2 with query-level examples whose observed retrieval state and baseline answer quality are similar but whose repair consequences differ sharply.

Table 7: Context composition for the coverage controls and their automatic baseline, averaged over 300 queries. All recall values divide distinct retained annotated facts by all annotated supporting facts.

Dataset	Condition	Units	Tokens	Annotation-nonsupport	Support recall	Truncation
2Wiki	AUTOMATIC	6.70	347.7	4.12	0.558	0.0%
2Wiki	ADD_ONLY	7.84	407.1	4.12	1.000	0.0%
2Wiki	FIXED_K_REPLACE	6.70	341.5	3.06	0.976	39.7%
2Wiki	REMOVE_ONLY	5.60	292.9	3.02	0.558	0.0%
HotpotQA	AUTOMATIC	6.64	363.8	4.30	0.604	0.0%
HotpotQA	ADD_ONLY	7.63	418.6	4.30	1.000	0.0%
HotpotQA	FIXED_K_REPLACE	6.64	358.4	3.38	0.974	33.0%
HotpotQA	REMOVE_ONLY	5.67	313.5	3.33	0.604	0.0%

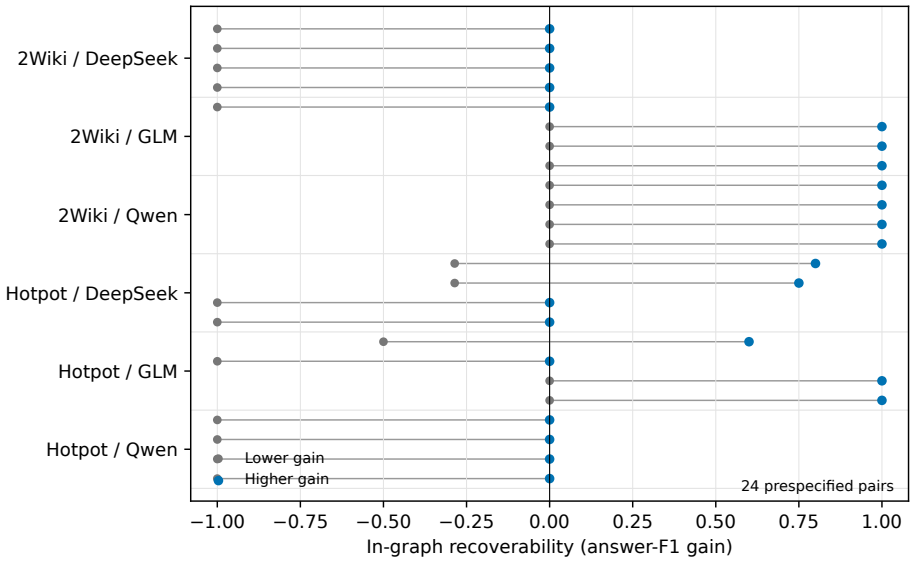

Figure 6: Metric-matched discordant cases grouped by dataset and generator. The four largest qualifying pairs per setting differ by at most 0.05 in support recall and 0.10 in baseline F1, yet by at least 0.50 in retrieval recoverability. The display is a prespecified diagnostic illustration rather than a representative sample of all queries.

B.5 Coverage-vs.-Noise Controls

ADD ONLY preserves every automatic unit and appends missing annotated support. REMOVE ONLY adds no support and removes the lowest-ranked annotation-nonsupport subset whose canonical-token mass is closest to the added mass. FIXED-K REPLACE substitutes missing support for annotation-nonsupport units one-for-one, preserves the automatic unit count, and truncates source text only as needed to meet the automatic token ceiling. All conditions use the same serializer, prompt, and deterministic Qwen3-8B manifest.

Exact provenance audits find no supporting-fact loss under `REMOVE_ONLY`: automatic and pruned recall are 0.558 on 2Wiki and 0.604 on HotpotQA. Both use all annotated facts as the denominator. The constrained repair never exceeds the automatic unit count or token budget. Its constrained/full-oracle fraction is descriptive. The coverage controls likewise identify the dominant tested contributor without providing a complete causal decomposition.

C Cross-Architecture Check

We evaluate HippoRAG 2 version `2.0.0-alpha.4` as a distinct architecture with an unmodified core checkout and a corpus adapter that deduplicates passages and excludes gold labels. It uses

Table 8: Qwen3-8B coverage and context-composition controls. Effects are paired against the frozen automatic condition.

Dataset	Condition	Gain (95% CI)
2Wiki	ADD_ONLY	+0.198 [+0.152, +0.244]
2Wiki	REMOVE_ONLY	-0.006 [-0.024, +0.012]
2Wiki	FIXED_K_REPLACE	+0.168 [+0.124, +0.211]
HotpotQA	ADD_ONLY	+0.187 [+0.146, +0.229]
HotpotQA	REMOVE_ONLY	+0.007 [-0.008, +0.023]
HotpotQA	FIXED_K_REPLACE	+0.123 [+0.084, +0.163]

Table 9: Cross-architecture evidence-interface sensitivity with Qwen3-8B. The primary-pipeline row uses the full in-graph provenance-filter condition; the HippoRAG 2 row uses its 1,024-token matched condition. Entries are paired answer-F1 gains with 95% bootstrap intervals.

Dataset	Architecture	Selection / interface gain	Serialization
2Wiki	Primary (in-graph filter)	+0.293 [+0.238, +0.347]	-0.004 [-0.028, +0.019]
2Wiki	HippoRAG 2 (token-matched)	+0.136 [+0.091, +0.183]	+0.020 [-0.011, +0.052]
HotpotQA	Primary (in-graph filter)	+0.236 [+0.189, +0.287]	-0.027 [-0.048, -0.008]
HotpotQA	HippoRAG 2 (token-matched)	+0.080 [+0.044, +0.117]	+0.005 [-0.013, +0.024]

normalized `BAAI/bge-small-en-v1.5` embeddings, DeepSeek OpenIE with a prompt-hash cache, native ranking depth 200 and top-$k = 5$, and a 1,024-token matched condition without gold-based truncation. Offline indexing contains 2,078 passages and 19,860 valid triples for 2Wiki and 2,976 passages and 28,145 triples for HotpotQA. We omit construction because no architecture-preserving construction oracle is defined for this system.

D Implementation and Statistical Details

D.1 Generator Manifests

DeepSeek uses `deepseek-v4-flash` with thinking disabled, temperature 0, top-$p = 1$, and 128 output tokens. GLM uses `glm-5.2` with thinking disabled, temperature 1.0, top-$p = 0.95$, and 128 output tokens. Confirmatory cells use three fresh calls with requested seeds 2026081701, 2026081702, and 2026081703. Each record stores the requested and returned model, decoding parameters, requested seed, context and response hashes, token usage, latency, timestamp, and retry status. The provider accepted the requested seed field for 100% of recorded DeepSeek and GLM calls; this records request acceptance, not proprietary-backend determinism or immunity to model-version drift. Qwen uses the official `Qwen3-8B-Q4_K_M.gguf` served by llama.cpp on Vulkan with thinking disabled, temperature 0, 128 output tokens, a 32,768-token context, and seed 42. The edge-count-aligned confirmatory control is reported only for DeepSeek and GLM, for which both matched construction cells are regenerated under the same prompt, decoding, and seed-pairing protocol. Comparisons are paired within generator and backend; absolute F1 differences between generator families are not evidence for recoverability.

D.2 Statistical Protocol

Paired contrasts are formed within query and generation replicate before inference. Primary stage-repair intervals use 10,000 paired query-bootstrap resamples. Original and protocol-controlled construction–retrieval confirmations and their outcome-scale sensitivity analysis use the unified 10,000-resample implementation in Appendix G; historical fixed-search intervals use 100,000. Hierarchical resamples first sample queries with replacement and then sample the three paired generation replicates within each selected query. Historical fixed-search sign-flip tests resample within-query generations before applying query-cluster signs. The revised interaction and new extension tests apply signs to the observed three-generation query means (100,000 permutations, plus-one correction). The fixed-search primary family contains 12 repair(k) minus automatic(16) contrasts for $k = 4, 8, 12$; the

Table 10: Answer-F1 cell means for the complete 2^3 factorial. Brackets report 95% paired query-bootstrap intervals.

Cell (C, R, S)	DeepSeek	GLM
$(0, 0, 0)$	$0.447\ [0.393, 0.501]$	$0.445\ [0.392, 0.499]$
$(1, 0, 0)$	$0.952\ [0.927, 0.973]$	$0.876\ [0.840, 0.910]$
$(0, 1, 0)$	$0.786\ [0.746, 0.827]$	$0.799\ [0.759, 0.837]$
$(0, 0, 1)$	$0.465\ [0.410, 0.519]$	$0.384\ [0.333, 0.436]$
$(1, 1, 0)$	$0.965\ [0.943, 0.983]$	$0.955\ [0.932, 0.975]$
$(1, 0, 1)$	$0.938\ [0.910, 0.963]$	$0.936\ [0.909, 0.961]$
$(0, 1, 1)$	$0.767\ [0.725, 0.809]$	$0.772\ [0.729, 0.813]$
$(1, 1, 1)$	$0.952\ [0.927, 0.973]$	$0.948\ [0.924, 0.970]$

secondary family contains 16 same-k contrasts. Holm correction is applied separately to each family and outcome. The complete original operating-point sensitivity family uses one eight-way correction. Exact match uses normalized string equality; the logit analysis transforms each bootstrap cell mean after a half-count correction; the headroom-normalized interaction divides the additive F1 interaction by the automatic cell's remaining population headroom. All other stratified and mechanism analyses are exploratory.

Answer F1 follows the released implementation. Both strings are lowercased; ASCII punctuation and the English articles `a`, `an`, and `the` are removed; whitespace is collapsed; and the result is split on spaces. F1 is the harmonic mean of precision and recall computed from multiset token overlap, with the usual equal-empty-string special case.

Artifact package. The accompanying archive contains ordered query IDs and cohort hashes, complete prompts, sanitized API manifests, condition and context hashes, raw per-query/per-replicate outputs, analysis scripts, tests, and environment lock files. API keys are not stored in run artifacts.

E Recoverability-Surface Robustness and Exposure Details

E.1 Factorial Estimand and Alignment

For an intervened stage set $A \subseteq \{C, R, S\}$, the higher-order contrasts use the standard Möbius expression

$$I_A(x; e, \rho) = \sum_{B \subseteq A} (-1)^{|A|-|B|} Q_x(B, e; \rho), \tag{8}$$

with the outcome, operating point e, and intervention protocol ρ defined in Equation 5. The construction–retrieval contrast in Equation 4 is the corresponding two-stage term with automatic ordering.

The 2Wiki experiment evaluates all 2^3 combinations of automatic or repaired construction, retrieval, and serialization. Each generator uses the same ordered 300 query IDs in every condition. Construction chooses $G_c \in \{G_{\text{auto}}, G_{\text{oracle}}\}$. Retrieval then applies either default R8 automatic search $(2, 2, 8)$ or the uncapped gold-provenance filter to the selected graph; serialization changes only the deterministic evidence order. The main paper reports the construction-conditioned retrieval-operator Möbius coefficient from the four $S = 0$ cells, without averaging across generators. Because the provenance filter's opportunity set changes with c, this is not an interaction under a fixed candidate space.

The complete eight-cell table reports the frozen original one-generation factorial, every treatment-coded Möbius term, and a paired query-bootstrap interval. Main-text C–R estimates use the later three-generation confirmation of the four $S = 0$ cells. Its higher-order serialization terms are conditional on the original unmatched protocol. The complete factorial is retained to show all cell means and serialization terms, not to replace the repeated-generation estimate.

E.2 Outcome-Scale Sensitivity

All four prespecified transformations retain a negative interaction, with intervals excluding zero for both generators. Their magnitudes are not directly comparable: the logit interaction is non-collapsible and headroom normalization changes the unit. We therefore use these analyses as directional sensitivity evidence.

Table 11: Treatment-coded Möbius decomposition of the complete 2^3 answer-F1 surface. Brackets report 95% paired query-bootstrap intervals.

Term	DeepSeek	GLM
Baseline $\mu_\emptyset$	$0.447\ [0.393, 0.500]$	$0.445\ [0.391, 0.499]$
μ_C	$+0.505\ [+0.448, +0.562]$	$+0.432\ [+0.368, +0.495]$
μ_R	$+0.339\ [+0.281, +0.395]$	$+0.354\ [+0.303, +0.408]$
μ_S	$+0.018\ [-0.006, +0.042]$	$-0.061\ [-0.109, -0.011]$
μ_{CR}	$-0.326\ [-0.386, -0.266]$	$-0.276\ [-0.338, -0.213]$
μ_{CS}	$-0.031\ [-0.063, +0.000]$	$+0.120\ [+0.055, +0.184]$
μ_{RS}	$-0.037\ [-0.073, -0.002]$	$+0.033\ [-0.021, +0.086]$
μ_{CRS}	$+0.037\ [-0.003, +0.080]$	$-0.099\ [-0.166, -0.033]$

Table 12: Historical interactions across outcome scales with unified mean-of-three paired hierarchical intervals. Headroom normalization preserves the sign and cannot rule out score saturation.

Generator	Outcome scale	Interaction [95% CI]
DeepSeek	Answer F1 (additive)	$-0.321\ [-0.381, -0.260]$
DeepSeek	Answer F1 (logit cell means)	$-1.162\ [-1.639, -0.623]$
DeepSeek	Answer F1 (headroom-normalized)	$-0.573\ [-0.654, -0.486]$
DeepSeek	Exact match (additive)	$-0.279\ [-0.340, -0.217]$
GLM-5.2	Answer F1 (additive)	$-0.340\ [-0.391, -0.287]$
GLM-5.2	Answer F1 (logit cell means)	$-1.049\ [-1.494, -0.554]$
GLM-5.2	Answer F1 (headroom-normalized)	$-0.577\ [-0.645, -0.506]$
GLM-5.2	Exact match (additive)	$-0.299\ [-0.357, -0.242]$

The current HotpotQA graph representation has no architecture-preserving construction oracle. We omit the HotpotQA construction interaction rather than approximate it.

E.3 Original Retrieval Operating-Point Sweep

R4/R12/R16 use $(\text{seeds}, \text{hops}, \text{max edges})$ configurations $(2, 2, 4)$, $(3, 2, 12)$, and $(4, 3, 16)$. Dense, BM25, and connectivity weights remain fixed at $.55/.25/.20$. Because seeds, hops, and maximum edges change together, these are operating points rather than a single-factor capacity intervention. They are also distinct from the factorial default R8=$(2, 2, 8)$. We retain the table as a sensitivity analysis; the main claim uses prefixes of frozen R16 source lists.

Table 13: Complete repeated-generation operating-point results. Every cell contains 300 queries and three generations. Gain is in-graph repair F1 minus automatic F1; intervals are hierarchical query–generation 95% CIs.

Dataset	Generator	Operating point	Tokens	Automatic F1	Repair F1	Gain [95% CI]
2Wiki	DeepSeek	R4	196.10	0.400	0.604	0.204 [0.151, 0.257]
2Wiki	DeepSeek	R12	480.34	0.554	0.714	0.160 [0.107, 0.214]
2Wiki	DeepSeek	R16	610.78	0.594	0.741	0.147 [0.101, 0.195]
2Wiki	GLM-5.2	R4	196.10	0.335	0.567	0.233 [0.181, 0.282]
2Wiki	GLM-5.2	R12	480.34	0.505	0.728	0.222 [0.173, 0.274]
2Wiki	GLM-5.2	R16	610.78	0.571	0.741	0.170 [0.129, 0.213]
HotpotQA	DeepSeek	R4	211.75	0.544	0.758	0.215 [0.170, 0.262]
HotpotQA	DeepSeek	R12	535.25	0.629	0.792	0.163 [0.122, 0.205]
HotpotQA	DeepSeek	R16	704.55	0.649	0.779	0.130 [0.092, 0.168]
HotpotQA	GLM-5.2	R4	211.75	0.515	0.725	0.210 [0.161, 0.261]
HotpotQA	GLM-5.2	R12	535.25	0.614	0.782	0.168 [0.124, 0.214]
HotpotQA	GLM-5.2	R16	704.55	0.646	0.791	0.145 [0.107, 0.187]

The original cross-operating-point conclusion is sensitive to which lower point is chosen. The complete eight-member family is:

Table 14: Complete original operating-point family: repair at R4 or R12 minus automatic R16. Holm correction is over all eight rows.

Dataset	Generator	Repair point	F1 difference [95% CI]	Holm p
2Wiki	DeepSeek	R4	$+0.010\,[-0.044, +0.064]$	1.000
2Wiki	DeepSeek	R12	$+0.120\,[+0.072, +0.169]$	$< 10^{-4}$
2Wiki	GLM-5.2	R4	$-0.003\,[-0.053, +0.046]$	1.000
2Wiki	GLM-5.2	R12	$+0.157\,[+0.116, +0.200]$	$< 10^{-4}$
HotpotQA	DeepSeek	R4	$+0.109\,[+0.071, +0.149]$	$< 10^{-4}$
HotpotQA	DeepSeek	R12	$+0.143\,[+0.105, +0.183]$	$< 10^{-4}$
HotpotQA	GLM-5.2	R4	$+0.080\,[+0.037, +0.123]$	$< 10^{-3}$
HotpotQA	GLM-5.2	R12	$+0.137\,[+0.099, +0.175]$	$< 10^{-4}$

R4 is unresolved for both 2Wiki generators, whereas every R12 contrast is positive and significant. We therefore do not select a capacity value for the main claim.

E.4 Fixed-Search Same-Cap Gains

The frozen automatic graph with ordered R16 automatic and repaired source lists yields the following F1 gains for repair(k) minus automatic(k). Here k is only the final serialization cap. All 16 intervals exclude zero, and all Holm-adjusted p-values are below 2×10^{-4}.

Table 15: Same-cap gains under fixed R16 search, $\mathrm{repair}(k) - \mathrm{automatic}(k)$. Brackets report 95% paired hierarchical-bootstrap intervals; p_{H} is Holm-adjusted over the 16 contrasts for each outcome. Identical displayed $p_H = 1.60 \times 10^{-4}$ values reflect the plus-one Monte Carlo resolution and Holm correction with 100,000 draws.

Dataset	Generator	k	Answer F1 gain [95% CI]; p_{H}	EM gain [95% CI]; p_{H}
2Wiki	DeepSeek	4	$+0.218\,[+0.168, +0.268]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.183\,[+0.131, +0.236]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$
2Wiki	DeepSeek	8	$+0.243\,[+0.192, +0.294]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.209\,[+0.157, +0.262]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$
2Wiki	DeepSeek	12	$+0.173\,[+0.124, +0.223]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.154\,[+0.103, +0.207]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$
2Wiki	DeepSeek	16	$+0.147\,[+0.101, +0.194]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.129\,[+0.081, +0.178]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$
2Wiki	GLM	4	$+0.246\,[+0.201, +0.292]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.206\,[+0.160, +0.252]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$
2Wiki	GLM	8	$+0.257\,[+0.211, +0.303]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.219\,[+0.172, +0.267]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$
2Wiki	GLM	12	$+0.209\,[+0.168, +0.251]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.189\,[+0.146, +0.233]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$
2Wiki	GLM	16	$+0.170\,[+0.132, +0.210]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.158\,[+0.117, +0.200]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$
HotpotQA	DeepSeek	4	$+0.198\,[+0.155, +0.242]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.174\,[+0.130, +0.221]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$
HotpotQA	DeepSeek	8	$+0.167\,[+0.126, +0.209]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.154\,[+0.111, +0.200]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$
HotpotQA	DeepSeek	12	$+0.136\,[+0.100, +0.174]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.129\,[+0.090, +0.170]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$
HotpotQA	DeepSeek	16	$+0.130\,[+0.093, +0.168]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.121\,[+0.083, +0.161]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$
HotpotQA	GLM	4	$+0.210\,[+0.167, +0.254]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.181\,[+0.137, +0.228]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$
HotpotQA	GLM	8	$+0.178\,[+0.138, +0.221]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.170\,[+0.126, +0.216]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$
HotpotQA	GLM	12	$+0.148\,[+0.110, +0.187]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.136\,[+0.094, +0.178]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$
HotpotQA	GLM	16	$+0.145\,[+0.109, +0.183]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$	$+0.136\,[+0.096, +0.178]$; $p_{\mathrm{H}} = 1.60 \times 10^{-4}$

For the cross-cap family repair(k) minus automatic(16), Table 16 reports F1 and exact match. Eleven of the 12 contrasts are positive after Holm correction in both metrics; the exception is 2Wiki DeepSeek at $k = 4$.

Table 16: Answer-F1 and exact-match differences between a targeted prefix and the automatic $k = 16$ prefix under fixed R16 search. Each contrast pairs 300 queries across three generations. Holm p-values are corrected over the 12-contrast family; † denotes $p \geq .05$.

Dataset	k	DeepSeek F1 Δ [95% CI]	DeepSeek EM Δ [95% CI]	GLM-5.2 F1 Δ [95% CI]	GLM-5.2 EM Δ [95% CI]
2Wiki	4	$+0.039\,[-0.015, +0.093]^{\dagger}$	$+0.034\,[-0.020, +0.090]^{\dagger}$	$+0.071\,[+0.021, +0.120]$	$+0.074\,[+0.023, +0.126]$
2Wiki	8	$+0.135\,[+0.088, +0.182]$	$+0.118\,[+0.069, +0.168]$	$+0.156\,[+0.114, +0.198]$	$+0.142\,[+0.099, +0.187]$
2Wiki	12	$+0.144\,[+0.097, +0.192]$	$+0.129\,[+0.080, +0.179]$	$+0.177\,[+0.139, +0.217]$	$+0.163\,[+0.122, +0.206]$
HotpotQA	4	$+0.130\,[+0.092, +0.169]$	$+0.122\,[+0.084, +0.162]$	$+0.129\,[+0.088, +0.170]$	$+0.118\,[+0.076, +0.161]$
HotpotQA	8	$+0.134\,[+0.096, +0.172]$	$+0.121\,[+0.082, +0.162]$	$+0.142\,[+0.106, +0.181]$	$+0.132\,[+0.091, +0.176]$
HotpotQA	12	$+0.136\,[+0.100, +0.175]$	$+0.127\,[+0.089, +0.167]$	$+0.141\,[+0.105, +0.179]$	$+0.129\,[+0.089, +0.170]$

E.5 Relation-Only and Coverage Controls

The Qwen3-8B relation-only serializer reuses the same selected edge IDs and performs no reretrieval. It removes source sentences, titles, sentence indices, and provenance metadata, leaving only extracted subject–relation–object triples. The gains are $+0.230$ (95% CI $[+0.181, +0.279]$) on 2Wiki and $+0.158$ ($[+0.112, +0.207]$) on

HotpotQA; relative to provenance-carrying gains, the attenuations are -0.051 and -0.056. ADD ONLY retains automatic units and appends missing annotated support; REMOVE ONLY deletes a similarly sized low-ranked annotation-nonsupport subset; FIXED-K REPLACE substitutes missing support under the automatic count and token ceilings. Their 2Wiki/HotpotQA gains are $+0.198/+0.187$, $-0.006/+0.007$, and $+0.168/+0.123$, respectively. These controls support an evidence-selection interpretation without providing a complete causal decomposition.

F SCOPE OF THE DIAGNOSTIC COMPARISON

Table 17 summarizes the targets of related evaluation and attribution families. These distinctions concern the reported quantity rather than a hierarchy of methods. In particular, evidence-conditioned studies already use controlled comparisons and privileged references. Our focus is on specified GraphRAG stage replacements and how their measured responses depend on the candidate graph and evidence exposure.

The factorial contrasts in INTERVENERAG summarize differences among prespecified intervention cells. They do not by themselves allocate intrinsic component credit, establish a variance decomposition over an input distribution, or identify direct and indirect causal effects. Likewise, an annotation-guided replacement is an empirical diagnostic condition, not a proof of optimality. The intervention definitions and controls in the main paper determine the interpretation of each reported gain.

Table 17: Targets of related diagnostic and attribution families. INTERVENERAG measures final-answer responses to specified GraphRAG stage replacements.

Family	What is changed or observed	Typical output	Difference from INTERVENERAG
Component RAG metrics (Es et al., 2024; Saad-Falcon et al., 2024; Ru et al., 2024)	Retrieved context and generated answer	Recall, relevance, faithfulness, answer quality	Characterize retrieval and generation; can accompany intervention measurements.
Evidence-conditioned diagnosis (Xia, 2026; Zhang et al., 2026; Deka & Singh, 2026)	Compare retrieved, gold-reference, or other evidence conditions	Answer changes and evidence-utilization diagnostics	Closely related perspective; INTERVENERAG specifies GraphRAG stage replacements and tests their combinations.
Context attribution (Li et al., 2026)	Ablate individual context sentences	Changes in response distributions	Attributes responses to context; INTERVENERAG measures answer-quality changes after stage replacements.
Variance-based sensitivity (Sobol', 1993)	Vary inputs under a specified distribution	Contributions to output variance	Targets variance decomposition; INTERVENERAG summarizes prespecified intervention cells.
Shapley attribution (Shapley, 1953; Lundberg & Lee, 2017)	Evaluate feature or component coalitions	Average marginal contribution	Allocates credit; INTERVENERAG reports prespecified repair responses.
Causal mediation (Imai et al., 2010; Pearl, 2001)	Change treatment and analyze a mediator	Direct and indirect effects	Requires causal assumptions; INTERVENERAG makes protocol-qualified diagnostic claims.
InterveneRAG	Replace specified GraphRAG stages with fixed downstream operators	Paired final-answer changes and intervention interactions	Interprets measured responses under explicit candidate-graph and evidence-budget conditions.

G PROTOCOL AUDITS AND ADDITIONAL REPORTING

Ordering algorithm. Starting from the selected input-edge list, choose the edge with the largest sum of subject/question and object/question lexical overlap (lowercase alphanumeric tokens); break ties by input position. Initialize the reached-entity set with its endpoints. Repeatedly append the first remaining edge in input order touching any reached entity, or the first remaining edge if none touches the reached set, and add its endpoints. This deterministic connected-first ordering is neither an optimal oracle nor standard BFS/DFS. Evidence identity and content are held fixed.

Generator variation. Across 1,200 query–condition groups in the original three-generation C–R confirmation, answer strings vary in 2.92% of DeepSeek and 29.42% of GLM groups; F1 varies in 0.67% and 15.00%. Mean within-group F1 standard deviations are 0.00337 and 0.06311. These are empirical properties of the recorded calls, not guarantees about provider determinism. The distinct decoding settings are fixed within each generator's paired comparison; their stochastic variability is not assumed equal across generators.

Evidence units and small prefixes. In the 300-query 2Wiki cohort, 232 questions have two distinct annotated facts, two have three, and 66 have four. The complete oracle's mean 4.64 graph edges is not the number of distinct required facts. R16 repaired prefixes at $k = 4$ contain a mean 1.387 repeated-source units and fully cover annotations for 61% of queries. This is a descriptive limitation of the frozen source list, not a proof of

Table 18: Experimental overview. All comparisons use frozen query cohorts and are paired within query.

Question	Data and generators	Controlled change	Main measurement
Can component metrics locate recoverability?	2WikiMultiHopQA and HotpotQA; DeepSeek, Qwen3-8B, GLM-5.2	None; predict in-graph repair gain from five automatic statistics	Query-disjoint cross-validated R^2
How much headroom remains in the graph?	Both datasets; all three generators	Annotation-guided selection inside the frozen automatic graph, with no larger unit or token budget	Paired answer-F1 gain
Do construction and retrieval repairs overlap?	2Wiki; DeepSeek and GLM confirmations	Historical factorial and exposure controls; fresh eight-cell bounded replication	Construction–retrieval interaction
Does selection beat more context?	Both datasets; DeepSeek and GLM	Fixed R16 search; automatic versus targeted prefixes at $k = 4, 8, 12, 16$	F1 and exact-match contrasts
Do selection gains transfer?	MuSiQue and shared HippoRAG indices; DeepSeek and GLM, three generations	Existing paragraph-graph units or indexed passages, native budget ceilings	F1/EM gain and support retention
Does order or prefix composition matter?	2Wiki; DeepSeek and GLM, three generations	Pure random/reverse/connected order; separate distinct-fact-first $k = 4$	Paired contrasts with Holm correction

the cause of a nonsignificant answer gain. Source-list repair prioritizes support edges by automatic rank, whereas the primary stage selector first covers distinct facts. For the source-deduplicated variant, deduplication reduces mean counts by 1.447/1.323 units (2Wiki/HotpotQA), and token trimming removes a further 0.057/0.077 units.

Table 19: Realized fixed-prefix evidence exposure. Each row summarizes 300 unique contexts before generator replication. These canonical counts measure evidence exposure, not total system cost.

Dataset	List	k	Mean tokens	Median tokens	Support recall
2Wiki	automatic	4	205.0	201.0	0.508
2Wiki	automatic	8	387.3	393.0	0.599
2Wiki	automatic	12	522.2	555.0	0.681
2Wiki	automatic	16	610.8	636.0	0.719
2Wiki	repair	4	209.8	204.0	0.805
2Wiki	repair	8	384.3	391.0	0.901
2Wiki	repair	12	509.7	549.0	0.921
2Wiki	repair	16	582.2	618.0	0.925
HotpotQA	automatic	4	225.4	220.0	0.605
HotpotQA	automatic	8	429.5	421.0	0.699
HotpotQA	automatic	12	587.0	611.5	0.741
HotpotQA	automatic	16	704.6	769.0	0.756
HotpotQA	repair	4	222.9	216.0	0.889
HotpotQA	repair	8	425.5	427.0	0.951
HotpotQA	repair	12	577.8	612.0	0.957
HotpotQA	repair	16	669.8	732.0	0.957

Extraction prompt (compact_v2). The system prompt reads:

> Extract factual relation triples from the supplied paragraph. Use only explicit facts in the paragraph. Do not infer, merge, or invent entities. Copy source_title and source_text exactly, and use the zero-based source_sentence_id. Return only one JSON object with this shape:
>
> `{"triples":[{"subject":"...","relation":"...","object":"...",`
> `"source_title":"...","source_sentence_id":0,`

Table 20: Absolute answer F1 for the primary support-first in-graph intervention (300 queries per setting).

Dataset	Generator	Automatic	Repaired	Gain
2Wiki	deepseek	0.465	0.739	+0.275
HotpotQA	deepseek	0.589	0.795	+0.206
2Wiki	qwen	0.387	0.668	+0.281
HotpotQA	qwen	0.515	0.730	+0.215
2Wiki	glm	0.422	0.761	+0.339
HotpotQA	glm	0.560	0.781	+0.221

Table 21: Historical HippoRAG 2 evidence-interface check with Qwen3-8B; automatic and intervened absolute F1. This table is not a graph-constrained selector replication.

Dataset	Automatic	Interface intervention	Gain
2Wiki	0.499	0.635	+0.136
HotpotQA	0.643	0.723	+0.080

```
"source_text":"exact sentence","confidence":0.0}]}
```

Use an empty triples array when there are no factual relations. Do not output markdown or explanations. Return at most one triple per source sentence and at most 12 triples total. Prefer the most informative named-entity relation in each sentence.

The user template is `Title: {title}`, then `Sentences:`, followed by one `[index] sentence` line per zero-based sentence and `Return the JSON object only.` Validation checks exact title, valid index, whitespace-normalized source text, nonempty relation fields, and confidence in $[0, 1]$.

Nonlinear prediction audit. Using the same five annotation-derived component features and frozen query-level gains, we fit a constant baseline, ridge, degree-two polynomial ridge, and histogram gradient boosting. Hyperparameters are selected using three inner folds within each of five outer query folds; preprocessing is fitted only on training folds. The target is never used to select outer test queries. Quadratic features reach $R^2 = 0.099$–0.146 and boosting 0.028–0.109, versus 0.091–0.138 for the ridge fit under this new split. This supports limited predictive power of these features under the tested models, not impossibility for all nonlinear predictors. The artifact also contains MAE, rank correlations, and 2,000 paired bootstrap intervals over fixed out-of-fold predictions; these intervals condition on the fitted folds and exclude retraining uncertainty.

Table 22: Component-only out-of-fold prediction of in-graph gain: nested five outer / three inner query folds shared by all models, seed 20260919. No baseline answer F1 is included. Cells are out-of-fold R^2; this is a separate fold/tuning protocol from the historical ridge result.

Dataset	Generator	Constant	Ridge	Degree-2 ridge	Gradient boosting
2Wiki	DeepSeek	-0.004	0.093	0.099	0.070
2Wiki	GLM-5.2	-0.011	0.137	0.142	0.109
2Wiki	Qwen3-8B	-0.012	0.091	0.104	0.072
HotpotQA	DeepSeek	-0.004	0.138	0.146	0.028
HotpotQA	GLM-5.2	-0.004	0.134	0.135	0.069
HotpotQA	Qwen3-8B	-0.001	0.096	0.099	0.038

Unified interaction uncertainty and cell means. Historical original, edge-count, and edge/token sensitivity tables now share one analysis implementation: 10,000 resamples, seed 20260919, queries sampled with replacement, and three paired replicate indices sampled with replacement within each sampled query. Earlier interaction summaries sampled one replicate per selected query, although their point estimates averaged three; the revised implementation targets the reported mean-of-three design. Point estimates are unchanged. Table 23 makes the four cells and the score-ceiling constraint explicit. The historical edge/token control constrains the selected $C_1 R_0$ context; it does not retroactively constrain the other historical cells.

Historical prediction and case selection. The earlier non-nested five-fold ridge analysis yields $R^2 = 0.064$–0.153 (median 0.114), distinct from the nested-CV split in the updated main figure. Adding baseline F1

Table 23: Four absolute F1 cell means for historical comparison protocols; all average three generations. The bound is $1 - Q_{10} - Q_{01} + Q_{00}$, not an independently estimated mechanism.

Generator	Protocol	Q_{00}	Q_{10}	Q_{01}	Q_{11}	I_{CR} upper bound
DeepSeek	Original (300)	0.440	0.945	0.776	0.959	-0.280
DeepSeek	Edge count (300)	0.440	0.685	0.776	0.935	-0.021
DeepSeek	Original (292)	0.439	0.943	0.772	0.958	-0.277
DeepSeek	Edge + token (292)	0.439	0.692	0.772	0.958	-0.025
GLM-5.2	Original (300)	0.412	0.930	0.781	0.960	-0.299
GLM-5.2	Edge count (300)	0.412	0.648	0.781	0.907	-0.017
GLM-5.2	Original (292)	0.409	0.928	0.783	0.959	-0.302
GLM-5.2	Edge + token (292)	0.409	0.651	0.783	0.959	-0.025

Table 24: Historical construction–retrieval protocol sensitivity on 300 queries. Intervals use the unified mean-of-three hierarchical bootstrap. Matching changes composition as well as full-graph edge count.

Generator	Original I_{CR}	Matched I_{CR}	Paired change
DeepSeek	-0.321 $[-0.381, -0.260]$	-0.086 $[-0.149, -0.025]$	$+0.235$ $[+0.182, +0.288]$
GLM-5.2	-0.340 $[-0.391, -0.287]$	-0.111 $[-0.166, -0.055]$	$+0.229$ $[+0.181, +0.278]$

gives $R^2 = 0.312$–0.493, but the target gain already subtracts this score; it is not a component-only predictor. Partial Spearman correlations between support recall and retrieval gain, controlling for baseline F1, range from -0.230 to -0.069. The illustrative case selection takes the four largest-gap qualifying pairs per setting: recall difference at most 0.05, baseline-F1 difference at most 0.10, and gain difference at least 0.50. These 24 selected pairs do not estimate prevalence. One 2Wiki–DeepSeek pair has recall 0.50 and baseline F1 1.0 for both queries, but gains of -1.0 and 0.0.

H Fresh Extensions and Fully Bounded Factorial

H.1 Frozen Protocol and Generation Records

The extension cohorts, contexts, model requests, and scoring rules are frozen before answer generation. All four studies use DeepSeek and GLM, with three fresh calls per condition (requested seeds 2026091901–2026091903). DeepSeek requests `deepseek-v4-flash` and returns `deepseek-flash`; GLM requests and returns `glm-5.2`. These calls are a separately timestamped model snapshot and are not pooled with historical generations. Temperatures/top-p are 0/1 and 1/0.95, respectively, with thinking disabled and a 256-token answer limit. Raw responses, requested/returned model identifiers, provider usage, context/prompt hashes, timestamps, retries, and decoding are retained. A provider accepting a seed does not guarantee deterministic sampling.

Both models receive exactly the same answer-system prompt as the historical pipeline and a shared question/evidence template. Supporting labels, decompositions, and answer strings are absent from generation messages. Explicit abstention with a null or missing answer is parsed as an empty answer, without retrying for another answer. Boolean answer values are preserved as literal JSON text (true/false), without inferring a yes/no answer; parser recovery is logged separately from API calls. API failures are retried and never scored as ordinary incorrect answers. Primary answer-text F1/EM matches the historical scoring convention; separate abstention fractions and abstention-zeroed F1 are supplied in the artifact. The 37,320 answer cells comprise 7,200 MuSiQue, 7,200 HippoRAG, 14,016 factorial, 3,504 additional ordering, and 5,400 small-prefix calls.

Table 25: Historical edge-count and selected-C1R0 token-envelope control on the predefined 292-query cohort. Other cells retain their historical definitions; this is not a fully bounded factorial. Intervals use the unified mean-of-three bootstrap.

Generator	Original I_{CR}	Matched I_{CR}	Paired change
DeepSeek	-0.319 $[-0.379, -0.257]$	-0.067 $[-0.132, -0.001]$	$+0.252$ $[+0.200, +0.303]$
GLM-5.2	-0.343 $[-0.396, -0.289]$	-0.066 $[-0.125, -0.006]$	$+0.277$ $[+0.229, +0.325]$

There are also 4,085 unique paragraph extraction tasks; failed attempts are separately recorded. These counts are workload descriptions, not claims of cost savings.

All new intervals use the shared 10,000-resample mean-of-three query/replicate bootstrap. Two-sided sign-flip tests use 100,000 query-level sign permutations. Holm families are separate for F1 and EM: two MuSiQue primary contrasts, four MuSiQue sensitivity contrasts, four shared-index primary contrasts, 14 factorial terms, six ordering contrasts, and six small-prefix contrasts. Hop-stratified primary gains are exploratory and receive their own six-test correction. Reported confidence intervals are marginal rather than simultaneous. No experiment or query is omitted for an unfavorable outcome.

H.2 MuSiQue with Paragraph-Level Supervision

MuSiQue targets connected multihop reasoning through question composition and filtering (Trivedi et al., 2022); it is not assumed immune to shortcuts. We sample 100 answerable questions each with two, three, and four decomposition steps (seed 20260919) from HippoRAG's published 1,000-question evaluation subset, which contains 518/316/166 such questions. This is an equal-hop 300-question sample, not a fresh sample of the complete official development set. The artifact records the source SHA-256 and every query ID. Each query uses its supplied candidate paragraphs.

Paragraph IDs, rather than possibly repeated titles, define support. Sentence segmentation is deterministic and annotation independent. Compact-v2 extraction is applied once per distinct paragraph prompt; validation checks title, sentence index, source text, and nonempty triple fields, then retains the first valid triple per sentence and at most 12 per paragraph. A truncated JSON extraction is retried with a larger output allowance; it is not silently treated as an empty graph. Sentence provenance is mapped back to the original paragraph ID. No decomposition is converted into artificial gold relations, and no construction oracle is claimed.

Automatic retrieval uses the same BGE/BM25/connectivity R8 rule. The intervention first selects one automatically ranked edge from each represented supporting paragraph, then remaining support edges and ranked fillers, subject to the automatic unit/token ceilings. Final-source truncation is recorded. A no-context arm tests evidence sensitivity under an instruction to abstain when evidence is insufficient; it is not an unrestricted closed-book capability test. A missing-support arm deterministically hash-selects one annotated supporting paragraph, removes all its edges, and repeats selection under the original ceilings. Alternative or redundant evidence can remain, so a correct answer in this arm is not by itself evidence of cheating. Paragraph presence, even with an untruncated sentence, does not establish representation of every necessary relation.

Table 26: Fresh musique replication: absolute answer F1, exact match, and abstention fraction, averaged over three generations per query.

Dataset	Generator	Condition	F1	EM	Abstention
MuSiQue	DeepSeek	automatic	0.306	0.236	0.339
MuSiQue	DeepSeek	missing_support	0.451	0.398	0.274
MuSiQue	DeepSeek	no_context	0.000	0.000	1.000
MuSiQue	DeepSeek	paragraph_repair	0.599	0.520	0.154
MuSiQue	GLM-5.2	automatic	0.302	0.249	0.500
MuSiQue	GLM-5.2	missing_support	0.474	0.413	0.384
MuSiQue	GLM-5.2	no_context	0.024	0.016	0.961
MuSiQue	GLM-5.2	paragraph_repair	0.607	0.510	0.210

Table 27: Fresh musique F1 contrasts. Marginal 95% intervals use 10,000 mean-of-three hierarchical resamples; Holm-adjusted sign-flip tests use query means and 100,000 permutations. Families are declared in the protocol manifest.

Dataset	Generator	Contrast	Difference [95% CI]	p_H
MuSiQue	DeepSeek	repair_minus_automatic	$+0.293$ $[+0.243, +0.345]$	2e-05
MuSiQue	DeepSeek	repair_minus_no_context	$+0.599$ $[+0.549, +0.648]$	4e-05
MuSiQue	DeepSeek	repair_minus_missing_support	$+0.148$ $[+0.100, +0.195]$	4e-05
MuSiQue	GLM-5.2	repair_minus_automatic	$+0.305$ $[+0.252, +0.358]$	2e-05
MuSiQue	GLM-5.2	repair_minus_no_context	$+0.583$ $[+0.532, +0.631]$	4e-05
MuSiQue	GLM-5.2	repair_minus_missing_support	$+0.133$ $[+0.081, +0.183]$	4e-05

Table 28: MuSiQue exploratory hop strata, 100 queries each. Overall results give equal weight to the three hop strata; they are not estimates for the natural hop distribution of the full benchmark.

Hops	Generator	Automatic F1	Repair F1	Gain [95% CI]
2	DeepSeek	0.452	0.691	$+0.239$ [$+0.154$, $+0.328$]
2	GLM-5.2	0.424	0.709	$+0.284$ [$+0.195$, $+0.376$]
3	DeepSeek	0.243	0.573	$+0.330$ [$+0.233$, $+0.426$]
3	GLM-5.2	0.262	0.572	$+0.310$ [$+0.214$, $+0.408$]
4	DeepSeek	0.223	0.533	$+0.310$ [$+0.229$, $+0.395$]
4	GLM-5.2	0.218	0.540	$+0.321$ [$+0.239$, $+0.408$]

H.3 Strict Selection from Shared HippoRAG Indices

We reuse the frozen HippoRAG 2 shared indices and native PPR retrieval for the same 300-query cohorts. The indices contain 2,078/2,976 distinct passages (2Wiki/HotpotQA). For every candidate, passage ID and exact title/text are verified against graph chunk vertices, chunk metadata, and the embedding store. Support mapping uses original paragraph occurrences and sentence annotations; the selector never reads answer strings or generates missing passages. Both comparison arms use fresh generation.

The automatic condition supplies native top-five passages. Repair prioritizes supporting passages already in the whole frozen index, uses native top-200 PPR ranks for ordering (passage-ID ties outside that list), fills from native rank order, and enforces the automatic passage-count and canonical-token ceilings. The final passage may be truncated; full supporting-sentence retention is audited separately from passage membership. This tests indexed passage selection rather than fact-edge retrieval. Because the shared corpus is formed from benchmark candidate pools and contains the mapped support, it tests bounded shared-corpus transfer rather than full-Wikipedia candidate acquisition. Corpus scale and architecture change together, so differences from the primary pipeline cannot identify a pure architecture effect. The earlier Qwen evidence-interface experiment remains separate (Table 21).

Table 29: Fresh hipporag replication: absolute answer F1, exact match, and abstention fraction, averaged over three generations per query.

Dataset	Generator	Condition	F1	EM	Abstention
2Wiki	DeepSeek	automatic	0.652	0.562	0.068
2Wiki	DeepSeek	indexed_repair	0.768	0.657	0.008
2Wiki	GLM-5.2	automatic	0.698	0.600	0.143
2Wiki	GLM-5.2	indexed_repair	0.783	0.672	0.044
HotpotQA	DeepSeek	automatic	0.683	0.516	0.058
HotpotQA	DeepSeek	indexed_repair	0.791	0.621	0.007
HotpotQA	GLM-5.2	automatic	0.703	0.523	0.087
HotpotQA	GLM-5.2	indexed_repair	0.795	0.609	0.008

Table 30: Fresh hipporag F1 contrasts. Marginal 95% intervals use 10,000 mean-of-three hierarchical resamples; Holm-adjusted sign-flip tests use query means and 100,000 permutations. Families are declared in the protocol manifest.

Dataset	Generator	Contrast	Difference [95% CI]	p_H
2Wiki	DeepSeek	repair_minus_automatic	$+0.117$ [$+0.080$, $+0.154$]	4e-05
2Wiki	GLM-5.2	repair_minus_automatic	$+0.086$ [$+0.056$, $+0.117$]	4e-05
HotpotQA	DeepSeek	repair_minus_automatic	$+0.109$ [$+0.077$, $+0.143$]	4e-05
HotpotQA	GLM-5.2	repair_minus_automatic	$+0.092$ [$+0.059$, $+0.126$]	4e-05

H.4 Full Eight-Cell Bounded Factorial and Pure Ordering

All eight $C \times R \times S$ cells are freshly generated on the predefined 292-query 2Wiki feasible cohort. The eight exclusions are inherited from the historical, outcome-blind selected-oracle-core token-feasibility audit. For $C = 1$, oracle relation edges are padded with original automatic edges to exactly match the full automatic

Table 31: Realized context composition. MuSiQue recall is support-paragraph presence and the last column requires at least one untruncated source sentence per support paragraph (not all necessary relations). HippoRAG columns both report exact full-support-sentence retention. For $k = 4$, recall is distinct annotated facts and the last column is the fraction of queries retaining every annotated fact. Tokens use the canonical counter.

Dataset	Study	Condition	Units	Tokens	Recall	Full-retention statistic
MuSiQue	musique	automatic	5.29	281.6	0.478	0.478
MuSiQue	musique	missing_support	4.87	265.2	0.625	0.601
MuSiQue	musique	no_context	0.00	5.0	0.000	0.000
MuSiQue	musique	paragraph_repair	4.88	265.1	0.941	0.901
2Wiki	hipporag	automatic	5.00	710.1	0.884	0.884
2Wiki	hipporag	indexed_repair	4.91	684.7	0.999	0.999
HotpotQA	hipporag	automatic	5.00	587.3	0.812	0.812
HotpotQA	hipporag	indexed_repair	4.99	575.4	0.939	0.939
2Wiki	k4	automatic_k4	4.00	205.0	0.508	0.180
2Wiki	k4	distinct_fact_first_k4	3.99	208.8	0.925	0.827
2Wiki	k4	original_repair_k4	3.99	209.8	0.805	0.610

graph's edge count. Automatic retrieval reuses frozen cosine scores and reranks BM25 in the current graph. The new $R = 1$ selector prioritizes distinct annotated supporting facts, then remaining support edges and ranked fillers. Every cell is capped at the original automatic evidence-unit count and token budget; these are upper bounds, not exact context-length equalities. Consequently the construction core need not be fully exposed to generation. This is a newly bounded estimand, distinct from the historical uncapped provenance-filter factorial.

After evidence selection and text truncation, $S = 1$ applies the connected-first ordering algorithm to the retained items. Assertions verify equality of the evidence-ID/text multiset and token count for every S pair. Random and reverse order controls use the exact C_0R_0 items; each query's random permutation has a fixed hash-derived seed. Thus these controls change order only, without changing which source content survives truncation. Factorial terms are baseline-anchored Möbius contrasts, not variance allocations or intrinsic causal contributions.

Table 32: Fresh factorial replication: absolute answer F1, exact match, and abstention fraction, averaged over three generations per query.

Dataset	Generator	Condition	F1	EM	Abstention
2Wiki	DeepSeek	C0R0S0	0.369	0.311	0.346
2Wiki	DeepSeek	C0R0S1	0.360	0.298	0.350
2Wiki	DeepSeek	C0R0_random	0.367	0.308	0.342
2Wiki	DeepSeek	C0R0_reverse	0.347	0.288	0.339
2Wiki	DeepSeek	C0R1S0	0.742	0.643	0.078
2Wiki	DeepSeek	C0R1S1	0.735	0.637	0.076
2Wiki	DeepSeek	C1R0S0	0.544	0.500	0.253
2Wiki	DeepSeek	C1R0S1	0.532	0.487	0.252
2Wiki	DeepSeek	C1R1S0	0.893	0.832	0.006
2Wiki	DeepSeek	C1R1S1	0.879	0.821	0.007
2Wiki	GLM-5.2	C0R0S0	0.381	0.317	0.576
2Wiki	GLM-5.2	C0R0S1	0.373	0.312	0.567
2Wiki	GLM-5.2	C0R0_random	0.379	0.313	0.539
2Wiki	GLM-5.2	C0R0_reverse	0.374	0.312	0.549
2Wiki	GLM-5.2	C0R1S0	0.764	0.662	0.105
2Wiki	GLM-5.2	C0R1S1	0.761	0.660	0.110
2Wiki	GLM-5.2	C1R0S0	0.522	0.477	0.467
2Wiki	GLM-5.2	C1R0S1	0.528	0.479	0.465
2Wiki	GLM-5.2	C1R1S0	0.906	0.848	0.016
2Wiki	GLM-5.2	C1R1S1	0.898	0.840	0.010

H.5 Distinct-Fact Priority at a Four-Unit Prefix

On all 300 frozen 2Wiki queries, we compare the automatic R16 prefix, original repaired R16 prefix, and a distinct-support-fact-first reordering of that same repaired R16 list before taking $k = 4$. Each arm uses

Table 33: Fresh factorial F1 contrasts. Marginal 95% intervals use 10,000 mean-of-three hierarchical resamples; Holm-adjusted sign-flip tests use query means and 100,000 permutations. Families are declared in the protocol manifest.

Dataset	Generator	Contrast	Difference [95% CI]	p_H
2Wiki	DeepSeek	C	$+0.175$ [$+0.119$, $+0.231$]	0.00014
2Wiki	DeepSeek	R	$+0.374$ [$+0.321$, $+0.426$]	0.00014
2Wiki	DeepSeek	S	-0.008 [-0.025, $+0.008$]	1
2Wiki	DeepSeek	CR	-0.024 [-0.084, $+0.036$]	1
2Wiki	DeepSeek	CS	-0.004 [-0.029, $+0.020$]	1
2Wiki	DeepSeek	RS	$+0.001$ [-0.024, $+0.026$]	1
2Wiki	DeepSeek	CRS	-0.004 [-0.041, $+0.033$]	1
2Wiki	DeepSeek	C0R0S1_minus_default	-0.008 [-0.025, $+0.008$]	1
2Wiki	DeepSeek	C0R0_reverse_minus_default	-0.021 [-0.046, $+0.002$]	0.437
2Wiki	DeepSeek	C0R0_random_minus_default	-0.002 [-0.025, $+0.020$]	1
2Wiki	GLM-5.2	C	$+0.141$ [$+0.083$, $+0.200$]	0.00014
2Wiki	GLM-5.2	R	$+0.383$ [$+0.335$, $+0.431$]	0.00014
2Wiki	GLM-5.2	S	-0.007 [-0.037, $+0.022$]	1
2Wiki	GLM-5.2	CR	$+0.002$ [-0.053, $+0.058$]	1
2Wiki	GLM-5.2	CS	$+0.013$ [-0.030, $+0.056$]	1
2Wiki	GLM-5.2	RS	$+0.005$ [-0.031, $+0.041$]	1
2Wiki	GLM-5.2	CRS	-0.019 [-0.066, $+0.029$]	1
2Wiki	GLM-5.2	C0R0S1_minus_default	-0.007 [-0.037, $+0.022$]	1
2Wiki	GLM-5.2	C0R0_reverse_minus_default	-0.007 [-0.040, $+0.025$]	1
2Wiki	GLM-5.2	C0R0_random_minus_default	-0.002 [-0.032, $+0.028$]	1

two generators and three fresh calls. This changes the selected content and is deliberately separate from the pure-order experiment. It matches the four-unit cap but not token counts; realized tokens and complete-support coverage are reported. The focus on 2Wiki follows the previously observed small-prefix exception, rather than selecting the dataset by the new result.

Table 34: Fresh k4 replication: absolute answer F1, exact match, and abstention fraction, averaged over three generations per query.

Dataset	Generator	Condition	F1	EM	Abstention
2Wiki	DeepSeek	automatic_k4	0.349	0.299	0.419
2Wiki	DeepSeek	distinct_fact_first_k4	0.742	0.639	0.078
2Wiki	DeepSeek	original_repair_k4	0.613	0.524	0.198
2Wiki	GLM-5.2	automatic_k4	0.343	0.293	0.607
2Wiki	GLM-5.2	distinct_fact_first_k4	0.755	0.648	0.137
2Wiki	GLM-5.2	original_repair_k4	0.616	0.523	0.300

Table 35: Fresh k4 F1 contrasts. Marginal 95% intervals use 10,000 mean-of-three hierarchical resamples; Holm-adjusted sign-flip tests use query means and 100,000 permutations. Families are declared in the protocol manifest.

Dataset	Generator	Contrast	Difference [95% CI]	p_H
2Wiki	DeepSeek	original_minus_automatic	$+0.264$ [$+0.213$, $+0.315$]	6e-05
2Wiki	DeepSeek	distinct_minus_automatic	$+0.393$ [$+0.342$, $+0.445$]	6e-05
2Wiki	DeepSeek	distinct_minus_original	$+0.129$ [$+0.090$, $+0.169$]	6e-05
2Wiki	GLM-5.2	original_minus_automatic	$+0.273$ [$+0.224$, $+0.325$]	6e-05
2Wiki	GLM-5.2	distinct_minus_automatic	$+0.412$ [$+0.362$, $+0.463$]	6e-05
2Wiki	GLM-5.2	distinct_minus_original	$+0.138$ [$+0.099$, $+0.178$]	6e-05