FOUND BUT FORGOTTEN: THE EVIDENCE ASSEMBLY GAP IN RETRIEVAL-AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A retrieval-augmented generation pipeline can expose every annotated passage needed for a question and still fail because the selected reader context does not preserve the complete set. We call this the *evidence assembly gap* and separate three stages: evidence is found in the candidate pool, preserved under a bounded passage budget, and used by the reader. We introduce **CAST: Compositional Assembly from Search Traces**, which searches from graph-frontier traces, fuses native and generated-search candidates, and jointly selects five intact passages. On 1,000 MuSiQue questions, CAST improves token F1 by 11.89 points over a frozen native baseline. Holding the candidate pool, selector model, and reader fixed, joint selection improves F1 by 6.10 points over pointwise selection; both paired 95% intervals exclude zero, and feasible support-set omissions fall by 55.1%. Increasing the number of passages passed to the reader does not uniformly improve the recorded pipelines, showing that search breadth and reader breadth are separate design choices. Under the tested reader contract, retrieval quality alone can therefore hide a downstream assembly bottleneck.

1 INTRODUCTION

Retrieval-augmented generation can fail after retrieval has succeeded. A candidate pool may contain every annotated passage needed for a question, yet the bounded reader context can omit one passage and break the evidence chain. Marginal recall can hide this loss: useful passages may rank well individually while the required set is never preserved together. We call this failure the **evidence assembly gap**.

The gap is directly visible on 1,000 MuSiQue questions. A shared candidate pool contains every annotated support passage, with enough room to retain them within a five-passage reader budget, for 854 questions. A selector that scores passages independently preserves all of that support for 522 questions; a selector that considers the passages jointly preserves it for 705. Across all 1,000 questions, token F1 rises from 54.39 to 60.48 with the same candidate pools, language-model checkpoint, reader, and passage budget.

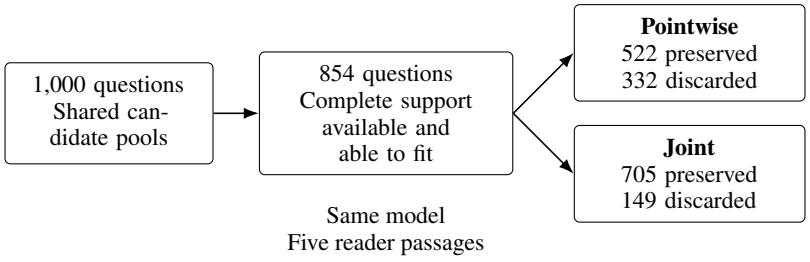

Figure 1: Joint selection preserves complete annotated support that pointwise selection discards, despite identical candidate opportunity. Both selectors have the same 854 feasible five-passage cases. "Preserved" means that every official support passage survives, not that semantic sufficiency has been independently established.

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.

Here, *forgotten* means that evidence found by search does not survive into the selected context; it does not imply that the language model has forgotten something it read. Figure 1 holds candidate opportunity fixed while selected evidence and answers change. The relevant unit is often a combination of passages: a bridge identifying an entity may be needed alongside a passage giving the requested fact.

We encountered this gap while trying to build the strongest controlled pipeline we could around an established graph-based retriever. We chose HippoRAG2 because it combines passage retrieval with relational structure, making it a useful foundation for questions whose answers require connected facts (Gutierrez et al., 2025). Our native implementation adds a bounded semantic walk and exposes the entities and bridge hypotheses reached during search. Freezing this HippoRAG2-derived core and the reader let us ask a narrower question: given a strong candidate opportunity and a fixed five-passage budget, how much answer quality is lost because the system assembles the wrong evidence set? The freeze isolates the surrounding search-and-selection pipeline; it does not claim that the resulting system is optimal.

CAST addresses this problem by separating how broadly the system searches from how much the reader consumes. It uses the native search trace to propose additional corpus-wide queries, combines their results with native candidates, and selects the reader's passages jointly. The method follows the principle *search wide, read narrow*, but does not assume that shorter contexts are always better. We test that question separately by varying the reader budget.

Contributions. We make three contributions. **(1)** We formulate and measure the assembly gap by separating evidence that is *found*, *preserved*, and *used*, and by counting annotated support that was available, able to fit, and nevertheless discarded. **(2)** We introduce CAST, a trace-conditioned method that combines corpus-wide discovery, duplicate-aware fusion, and joint fixed-budget assembly. **(3)** Under a frozen reader contract, we show controlled gains in support preservation and answer F1, while reader-budget and completion analyses identify limits to a simple "more evidence is better" explanation.

Scope. Evidence-set selection has substantial precedent (Lee et al., 2025; Meng et al., 2025; Sun et al., 2025; Jiang et al., 2026). Our experiments establish gains over the stated native and pointwise controls, not over those unrun published selectors. Official support sets are measurable proxies rather than independently validated minimal witnesses; the reader-relative oracle is conceptual; and the full candidate pool is never given to the reader. The detailed benchmark audit remains a secondary diagnosis of residual errors, not a claim that established QA tasks are solved.

The paper follows the same found–preserved–used sequence: Section 2 defines the seams, Section 3 describes CAST, and Sections 4–5 test discovery, assembly, and reader behavior.

2 From Found Evidence to Usable Answers

2.1 The unit of evidence is often a set

The assembly gap is easiest to see when an answer requires a bridge. Consider an illustrative, fictional question: *In which country was the author of The Glass Harbor born?* One passage identifies Mira Vale as the author; another states that Mira Vale was born in Iceland. The first resolves the bridge entity and the second supplies the requested property. Selecting several passages about the novel while omitting the biographical passage leaves the justification incomplete. This example illustrates the problem; it is not a benchmark observation.

To formalize the distinction, fix a question q, a finite corpus D, and a standard for a valid evidence-grounded answer. A *witness* is an inclusion-minimal passage set that justifies such an answer. If $\mathcal{W}_q$ is the family of valid witnesses, witness availability in context S is

$$A_q(S) = \mathbf{1}[\exists W \in \mathcal{W}_q : \; W \subseteq S] \,. \tag{1}$$

The definition has an AND–OR structure: all passages in one witness are needed together, but another complete witness may substitute for it. With witnesses $\{a, b\}$ and $\{c, d\}$, adding b is valuable when a is present; adding d to $\{a, b, c\}$ completes another witness but does not change whether any witness is available. This objective can exhibit both increasing and diminishing marginal returns.

That observation motivates joint decisions, not a claim that all ranking methods fail. Appendix A gives the formal argument and its assumptions.

2.2 Found versus preserved: an observable gap

Valid witnesses are generally unknown. A dataset's official support set H_q provides a measurable proxy, but may be nonminimal, miss alternatives, or contain errors. We therefore distinguish *complete annotated support* from *semantic sufficiency* throughout.

Let C_q be the effective candidate pool, S_q the selected subset, and k_q the passage budget. The pool records what search *found*; the selected subset records what assembly *preserved*. Each question belongs to one of four categories:

Category	Observable condition	What it locates
Absent	$H_q \not\subseteq C_q$	Search has not exposed all annotated support.
Over budget	$H_q \subseteq C_q$ and $\lvert H_q \rvert > k_q$	All support is present but cannot fit.
Feasible but discarded	$H_q \subseteq C_q$, $\lvert H_q \rvert \le k_q$, and $H_q \not\subseteq S_q$	Selection loses support that was available and could fit.
Preserved	$H_q \subseteq S_q$	All annotated support reaches the reader.

These categories partition the cohort for a valid exactly-k_q selection. The frequency of *feasible but discarded* cases is the observable **annotation-completeness gap**. Figure 1 measures this gap on a shared set of opportunities, rather than attributing every answer error to retrieval.

Why not use average support recall instead? Within questions with exactly h annotated supports, let r be mean support recall and c the fraction retaining the entire support set. Then

$$\max(0,\ hr - (h-1)) \le c \le r. \tag{2}$$

For four-support questions, recovering three supports on every question gives 75% recall and 0% completeness. Recovering all four on three-quarters of questions and none on the rest gives the same recall but 75% completeness. The bound is sharp without further constraints and requires no independence assumption. Completeness must therefore be measured directly; Appendix A supplies the proof.

2.3 Preserved versus used: a reader-relative gap

The second seam lies between selected evidence and the scored answer. Even a complete justification may be used poorly. Let $U_R(q, S; \tau)$ be expected answer utility for a fixed reader R, context S, and execution contract τ. The contract specifies ordering, prompting, output budget, answer extraction, and scoring. Utility is measured by that scorer, not by an independent semantic oracle.

For candidate pool C, budget k, and feasible selector π, define

$$\begin{aligned} V_R(q, C, k; \tau) &= \max_{\substack{S \subseteq C \\ \lvert S \rvert \le k}} U_R(q, S; \tau), \\ G_R &= V_R - U_R(q, \pi(q, C, k); \tau)\,. \end{aligned} \tag{3}$$

V_R is the best reader utility available from the pool; G_R is the opportunity the implemented selection fails to realize. With the contract fixed, a larger pool or allowed budget cannot reduce V_R: the old subsets remain available. Actual utility can nevertheless fall if selection changes, since $\Delta U_R = \Delta V_R - \Delta G_R$.

Equation (3) is a conceptual comparison, not an estimated regret or CAST's implemented objective. Its oracle permits *up to* k passages; CAST selects *exactly* k. The annotation-completeness gap is observable but is not the oracle utility gap. A reader can score well without all official support, or poorly despite retaining it. The reader-budget ablations in Section 5.3 change k and measure implemented utility without estimating V_R or G_R; the completion analysis changes our understanding of the execution contract without identifying a causal mediator. Together, these objects organize the experiments as *found*, *preserved*, and *used*.

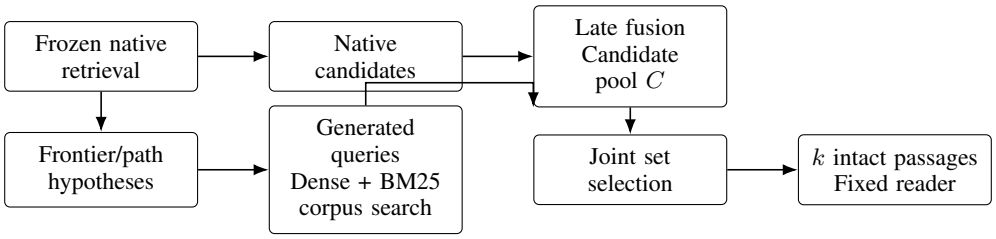

Figure 2: CAST searches broadly from graph-frontier traces, then jointly assembles exactly k intact passages for the fixed reader. Search traces suggest where to look; retrieved passages, not generated queries, provide the evidence.

3 CAST: Compositional Assembly from Search Traces

CAST maps the found–preserved–used decomposition onto three decisions: where to search next, which candidates to admit, and which passages should reach the reader together. It leaves the native retrieval trajectory and reader model unchanged.

3.1 Search from graph frontiers

The native procedure combines query-conditioned personalized PageRank with a bounded semantic walk over a HippoRAG2-derived graph. It returns an ordered candidate list N, an incumbent final five I, and frontier traces T. A frontier records an entity or bridge hypothesis reached by the walk, together with bounded path evidence. It is a lead for further search, not a verified conclusion.

For each selected frontier, the generator receives the question, the bridge hypothesis, and bounded excerpts of already retrieved text. It generates three distinct queries, each at most eight words **and** 96 characters. Each query searches the full benchmark corpus using dense retrieval and BM25, so new evidence need not lie within the bounded graph walk. For retrieval event e, document d receives reciprocal-rank-fusion score

$$s_e(d) = \frac{1}{60 + \mathrm{rank}_{e,\mathrm{dense}}(d)} + \frac{1}{60 + \mathrm{rank}_{e,\mathrm{BM25}}(d)}. \tag{4}$$

A missing rank contributes zero. Generated queries guide retrieval; they are not answers supplied to the reader.

3.2 Fuse without duplicate votes

The same query can be generated from different frontiers. Counting every occurrence as a new vote would inflate its documents' scores. For each distinct query string u, CAST keeps the best retrieval instance for d, then sums across distinct strings:

$$g_q(d) = \sum_{u \in \mathcal{Q}_q} \max_{e:\mathrm{query}(e)=u} s_e(d). \tag{5}$$

An absent document contributes zero. Ties are resolved by the number of distinct retrieving queries, best fused rank, first-seen order, and document identifier. Distinct frontier/query contexts remain available as provenance even when score contributions are deduplicated.

The proposed pool starts with the first a unique native candidates and appends at most b new generated candidates ranked by g_q. The headline setting is $a = 20$, $b = 15$, giving at most 35 candidates before packing. This protects native availability during admission; it does not oblige the final selector to retain native passages. The ablation does not establish an independent aggregate F1 gain from protection itself.

3.3 Assemble the reader context jointly

A single selector prompt sees the question, candidate passages, native path, source labels, and bounded retrieval provenance. It selects exactly k distinct labels according to the passages' com-

Algorithm 1 CAST: Compositional Assembly from Search Traces

Require: Question q; corpus D; frozen models and native retriever; admission caps $a = 20, b = 15$; selector ceiling L; reader $k = 5$.

Obtain native candidates N, incumbent I, and frontier trace T.
for each selected frontier f in T **do**
    Generate and validate three bounded queries from q, f, and path evidence.
    Retrieve full-corpus dense and BM25 results; retain ranks and provenance.
end for
Deduplicate documents and query contributions; rank new hits by g_q.
Concatenate the first a unique native candidates and at most b new hits.
Pack the selector input under L; record the effective pool C.
Jointly select k distinct labels from C.
if selection is invalid **then**
    Record the failure and use the declared fallback.
end if
Restore selected passages to pool order; render their intact text.
Run the fixed reader; extract the answer and record completion outcomes.

bined bridge and answer evidence. Generated queries and ranks are weak provenance, not proof. Selected passages remain intact and are restored to pool order before reading.

CAST separates the native walk budget, candidate-admission caps (a, b), selector input ceiling L, and reader passage budget k. Its headline reader sees five passages regardless of search breadth. Packing and fallback behavior are explicit parts of the evaluated method (Appendix B), not hidden repairs. CAST is a concrete composition of search and extractive selection; it neither enumerates witnesses nor certifies sufficiency or optimality in Equation (3).

4 A Controlled Test of the Assembly Gap

4.1 Controls isolate discovery, assembly, and reading

The design follows the same three stages as the problem formulation. The controlled study uses all 1,000 questions in the packaged MuSiQue cohort. Three controls distinguish the full intervention from its parts. The **native final-five baseline** uses the frozen retriever's own selection. **Pointwise selection** receives CAST's same expanded candidate pool and selector checkpoint, scores each passage independently from 0 to 9, and takes the top five with pool-position tie-breaking. **Native-only joint selection** applies the joint protocol to at most 20 native candidates.

Thus, the pointwise comparison holds the pool, model, reader, and passage budget fixed while changing the selection protocol. Native-only versus mixed-pool joint selection changes both source composition and candidate width. Reader-budget experiments separately vary k. These controls identify effects of the implemented interventions, not the causal contribution of every prompt feature.

Whole-pipeline evaluation uses every packaged question in MuSiQue (Trivedi et al., 2022), 2WikiMultiHopQA (Ho et al., 2020), HotpotQA (Yang et al., 2018), NaturalQuestions (Kwiatkowski et al., 2019), and PopQA (Mallen et al., 2023) (1,000 each), NarrativeQA (Kocisky et al., 2018) (293), and LV-Eval (Yuan et al., 2024) (124). These are not complete upstream splits. Native scores describe our frozen HippoRAG2-derived reproduction, not reported scores from the original HippoRAG2 paper. The earlier seven-cohort runs and the fresh MuSiQue controls are reported separately. For context, Table 4 also transcribes the published QA comparison on this cohort collection from HippoRAG 2. Those rows use a Llama-3.3-70B-Instruct reader, whereas our runs use Qwen3.6-27B; cross-block differences are therefore descriptive, not controlled effects.

4.2 Reader contract, measurements, and uncertainty

Selection and reading use the same Qwen3.6-27B snapshot with greedy decoding. Reader rendering, extraction, and input/output ceilings are held fixed within comparisons. The reader produces a rationale followed by an `Answer:` marker under a 1,024-token output ceiling; a missing or empty

Table 1: CAST improves F1 over the native final five, same-pool pointwise selection, and native-only joint selection. These are fresh paired MuSiQue comparisons on 1,000 questions with the reader and five-passage budget fixed. CAST's endpoint is one condition, not three independent replications; differences use unrounded per-question scores.

Baseline	Baseline F1	CAST F1	Paired gain [95% CI]
Native final five	48.59	60.48	$+11.89$ $[+9.37, +14.44]$
Pointwise selection, same mixed pool	54.39	60.48	$+6.10$ $[+3.73, +8.47]$
Joint selection, native cap 20	57.31	60.48	$+3.17$ $[+1.21, +5.14]$

Table 2: Joint selection preserves 705 of 854 feasible annotated support sets, versus 522 for pointwise selection on the identical mixed pool. The table reports candidate opportunity and preservation on the same 1,000 questions; all available official support sets fit these five-passage conditions. Appendix C gives the full partitions, including other budgets.

Candidate pool and selector	Mean width	Available and fitting	Feasible but discarded	Preserved
Native cap 10, joint 5	10.00	613	38	575
Native cap 20, joint 5	16.30	711	92	619
Mixed cap 35, pointwise 5	26.16	854	332	522
Mixed cap 35, joint 5	26.16	854	149	705

answer becomes `UNKNOWN`. This matters because better selected evidence may affect completion as well as reasoning. Appendix B specifies model, packing, fallback, and reuse contracts.

We report normalized exact match, token F1, support recall, full annotated support, and lexical answer coverage in selected titles or passages. Coverage is not sufficiency. NarrativeQA coverage is restricted to the question's source document; NarrativeQA and LV-Eval lack support labels.

The original ablation contrasts were declared before execution. Confidence intervals use 20,000 paired question-level percentile-bootstrap replicates with seed 20260814. Feasibility partitions, support-count strata, budget interaction, and completion accounting are exploratory analyses of frozen outputs. All intervals are nominal, unadjusted 95% intervals conditional on these runs, not uncertainty over new model families, corpora, or repeated inference. No outcome-driven dataset-specific tuning followed the headline protocol freeze.

5 Results: From Found to Preserved to Used

5.1 Does CAST improve answers with the same reader?

Yes. With the frozen reader and $k = 5$, CAST improves F1 against every tested control.

The native comparison measures the complete augmentation. The pointwise comparison shows that discovering more candidates is not the entire explanation: selection still changes answer quality when those candidates are identical. The native-only joint comparison shows a further gain from the mixed pool. The boundary is that these overlapping contrasts are controls, not independent replications or an additive decomposition of the 11.89-point gain.

5.2 Does joint selection preserve evidence that was already available?

Yes. For the 854 questions with all annotated support available and able to fit, joint selection preserves 705 complete sets, versus 522 for pointwise selection. Conditional preservation rises from 61.12% to 82.55%; feasible omissions fall from 332 to 149, a 55.1% net reduction, $(332-149)/332$.

Joint selection also improves full support by 18.30 points [15.70, 21.00], support recall by 9.29 [7.99, 10.59], and lexical coverage by 8.90 [6.70, 11.10]. The assembly gap is therefore directly visible in selected evidence, not inferred from F1 alone.

Opportunity and preservation must be read together. Native cap 10 preserves 575 of its 613 feasible sets, whereas the wider mixed pool preserves 705 in total. A narrow pool can look better conditionally by never exposing difficult opportunities. Broader native search also improves F1; Appendix C reports those width contrasts. Assembly does not replace discovery.

The gap remains substantial on difficult questions. For four-support questions, joint selection achieves 62.95% recall but only 18.67% complete recovery. Full strata appear in Appendix C.

The boundary is that the 55.1% value is an aggregate failure-count reduction, not a claim that exactly 183 cases were repaired with no new failures. The result supports measuring complete sets; it does not establish that complementarity alone causes the answer gain.

5.3 Ablation study: which parts of CAST matter?

Table 3 makes the ablation design explicit: each row states which variants are compared, what changes, and the resulting paired F1 effect.

Table 3: Ablation map on 1,000 MuSiQue questions. Each row names the variants, changed factor, and paired F1 effect. The reader model, prompt, decoding, and rendering are fixed; passage count changes only in the two k rows. "Bundled" marks a multi-factor comparison.

Ablation	Variants compared	What changes	F1 effect [95% CI]
Native breadth	Native joint 10 → 20	Native cap; joint selector and $k = 5$ fixed	$+3.17$ $[+1.54, +4.81]$
Generated search	Native 20 → protected mixed 35	Generated candidates and width (bundled)	$+3.17$ $[+1.21, +5.14]$
Joint assembly	Same-pool pointwise → joint	Selector only; mixed cap 35 and $k = 5$ fixed	$+6.10$ $[+3.73, +8.47]$
Source fusion	Generated-only → protected mixed 35	Adds native candidates at the same requested cap	$+0.49$ $[-1.46, +2.41]$
Native protection	Unprotected → protected mixed 35	Candidate-admission rule only	$+0.15$ $[-1.72, +2.01]$
Pool ceiling	Protected mixed 35 → 50	Requested cap; realized mean width 26.16 → 27.48	-0.22 $[-0.76, +0.29]$
Native k	$5 \to 8$ and $5 \to 10$	Number of native-ranked passages read	$+3.04$ $[+0.88, +5.23]$; $+6.21$ $[+3.94, +8.54]$
CAST k	$3 \to 5$ and $5 \to 8$	Number selected; joint selection reruns at each k	$+0.73$ $[-1.25, +2.72]$; -0.42 $[-2.38, +1.56]$

The clearest isolated mechanism is joint assembly. On the identical mixed cap-35 pool, changing only pointwise scoring to joint selection improves F1 by 6.10 points and full support by 18.30 points. Native breadth also helps from cap 10 to 20. The cap-20 to mixed cap-35 gain bundles generated access with width; generated-only versus fusion is unresolved on F1 but fusion adds 9.70 points of full support. Neither native protection nor cap 35 to 50 has a detected aggregate F1 benefit.

More native-ranked passages improve F1, whereas CAST's five-to-eight change is unresolved despite 5.70 points more full support. The direct paired comparison of the two five-to-eight changes is -3.45 F1 points $[-6.39, -0.54]$: adding three passages helps native ranking more than CAST in these recorded pipelines. Appendix Table C6 reports every reader-budget endpoint.

The source-and-width contrast is bundled, the budget comparison is exploratory, and the null protection and cap-50 intervals do not establish equivalence or saturation. **We do not test giving the reader the entire retrieved candidate pool**.

5.4 Does support preservation explain the whole answer gain?

No. Under the fixed rationale-first reader contract, answer completion also changes with selection. Cases where the required answer marker appears for only one selector contribute 4.57 of the 6.10-point F1 difference, or 75.03%, using unrounded values. Appendix C gives the full marker-transition accounting.

Among the 507 questions for which both selectors preserve every official support passage, joint-minus-pointwise F1 is -1.29 points. This rules out describing the observed result as a uniform answer improvement once annotated support is complete.

The boundary is mechanistic. Selected evidence may affect reasoning and completion together, and a missing marker can reflect truncation, formatting, or another execution failure. The both-complete group is post-treatment and does not hold evidence constant or identify a causal reader effect. CAST improves preservation and scored answers overall, but their causal relationship remains unresolved.

5.5 Do whole-pipeline gains extend beyond MuSiQue?

Table 4 situates the frozen whole-pipeline runs beside the same-cohort QA comparison reported by HippoRAG 2.

Table 4: QA F1 on the seven-cohort collection used by HippoRAG 2. Published rows reproduce the per-dataset values in Table 2 of HippoRAG 2 (Gutierrez et al., 2025), which uses Llama-3.3-70B-Instruct as the reader. Our frozen reproduction and CAST use Qwen3.6-27B on the same packaged cohorts. The source table's aggregate column is omitted so every displayed value has an explicit dataset identity. The two blocks provide context, not a controlled cross-reader ranking.

Method	NQ	PopQA	MuSiQue	2Wiki	HotpotQA	LV-Eval	NarrativeQA
Published in HippoRAG 2: Llama-3.3-70B-Instruct reader							
None	54.9	32.5	26.1	42.8	47.3	6.0	12.9
Contriever	58.9	53.1	31.3	41.9	62.3	8.1	19.7
BM25	59.0	49.9	28.8	51.2	63.4	5.9	18.3
GTR (T5-base)	59.9	56.2	34.6	52.8	62.8	7.1	19.9
GTE-Qwen2-7B-Instruct	62.0	56.3	40.9	60.0	71.0	7.1	21.3
GritLM-7B	61.3	55.8	44.8	60.6	73.3	9.8	23.9
NV-Embed-v2 (7B)	61.9	55.7	45.7	61.5	75.3	9.8	25.7
RAPTOR	50.7	56.2	28.9	52.1	69.5	5.0	21.4
GraphRAG	46.9	48.1	38.5	58.6	68.6	11.2	23.0
LightRAG	16.6	2.4	1.6	11.6	2.4	1.0	3.7
HippoRAG	55.3	55.9	35.1	71.8	63.5	8.4	16.3
HippoRAG 2 (published)	63.3	56.2	48.6	71.0	75.5	12.9	25.9
Our frozen runs: Qwen3.6-27B reader							
HippoRAG 2 (our reproduction)	61.29	60.73	48.70	73.20	76.99	12.53	26.51
CAST (ours)	**63.47**	**62.14**	**60.05**	**80.78**	**79.49**	**15.57**	**30.24**

Our HippoRAG 2 reproduction is lower than the published row on NaturalQuestions (61.29 versus 63.3) and LV-Eval (12.53 versus 12.9), and higher on the other five cohorts. Because the published table uses a different reader and construction protocol, these gaps do not diagnose reproduction fidelity. CAST is numerically above every published row on all seven cohorts, but the cross-reader table does not support a controlled state-of-the-art claim.

Within our matched reader contract, CAST improves mean F1 over our reproduction on every cohort. Appendix Table C5 gives the paired effects and intervals: six exclude zero and LV-Eval remains inconclusive. Appendix B reports sign tests and selection-stage resource counts.

6 Relation to Existing Work

RAG systems face three adjacent problems: what retrieval *finds*, what context construction *preserves*, and what the reader and scorer ultimately *use*. CAST acts at the first two stages and evaluates all three.

Finding candidate evidence. HippoRAG2 supplies the graph-retrieval foundation (Gutierrez et al., 2025). Long-context RAG studies how retrieval and larger reading budgets interact (Jin et al., 2024). CAST instead uses graph-frontier traces to widen search while holding the headline reader budget fixed.

Preserving evidence sets. SetR identifies a query's information requirements and selects passages collectively (Lee et al., 2025); DPS conditions on earlier choices (Meng et al., 2025); Evidence Tree

Search builds evidence paths (Sun et al., 2025); and OptiSet combines expansion with set refinement (Jiang et al., 2026). CAST couples a frozen graph trace to corpus-wide discovery, duplicate-aware admission, and joint assembly. Its controlled gains are over our native and pointwise controls; Table 4 provides broader but cross-reader context.

Using and evaluating selected context. Sufficient Context separates whether context supports an answer from whether a model succeeds (Joren et al., 2024). eRAG and RAGChecker diagnose downstream usefulness (Salemi & Zamani, 2024; Ru et al., 2024); Recall Is Not Enough and RECOMP study context packing and compression (Bala, 2026; Xu et al., 2023); answer-equivalence work separates reader behavior from scoring (Bulian et al., 2022; Kamalloo et al., 2023). We operationalize one seam: complete annotated support that was available, fit, and was still discarded. The contribution is the method and controlled stage-wise evidence, not the invention of set selection or a universal case for compact context.

7 Limitations and Remaining Bottlenecks

Comparison validity. The strongest controlled result concerns one packaged MuSiQue cohort and one model family. Pointwise ties, stronger published selectors, and matched-cost question-only expansion remain necessary controls. Table 4 uses different readers across blocks and is contextual, not a controlled leaderboard.

Mechanism and reader dependence. Frontier conditioning, source expansion, and candidate width remain partly bundled. Reader families, answer-first prompting, frozen-evidence controls, and validated bridge deletions are needed to test reader dependence and complementarity.

Measurement. Official supports are proxies, not validated minimal witnesses, and the at-most-k oracle is conceptual. Exploratory intervals are unadjusted; marker and both-complete groups are post-treatment, not causal mediators.

Generalization and audit validity. The seven-cohort whole-pipeline results broaden the evidence but do not reproduce the controlled assembly mechanism across datasets or reader families. Near-complete annotated support on 2WikiMultiHopQA and HotpotQA coexists with lower F1, so residual error need not mean failed search. The model-assisted audit in Appendix E is diagnostic: it lacks human validation and false-acceptance checks on full-F1 outputs. Neither these labels nor near-complete support establish a semantic performance ceiling.

8 Conclusion

Evidence can be found without being preserved, and preserved without being used successfully. CAST uses search traces to broaden discovery, then jointly assembles five reader passages. With the reader fixed, it improves complete-support preservation and answer F1; the ablations identify joint assembly as the clearest isolated mechanism and show that additional reader context has pipeline-dependent returns.

The architectural implication is to control search breadth and reader breadth separately: expose alternatives broadly, preserve coherent evidence under budget, and evaluate what is found, lost during assembly, and ultimately used.

Ethics Statement

Public artifacts contain aggregate results, categorical audit labels, source code, and hashed provenance, not raw benchmark questions, passages, predictions, generated queries, or document identities. Reassessing private traces requires authorized access to the original corpora and reports. The illustrative example in Section 2 is fictional. Benchmark judgments should not be used to relabel or retire items without independent review.

AI Use Statement

AI assistants supported literature investigation, manuscript development, analysis implementation, and review. They assisted with structural critique, comparison against recent award papers, prose revision for Versions 8 and 9, transcription of the cited HippoRAG 2 comparison table, and conversion to modular LaTeX. The frozen model-assisted annotation protocol, including model identities and quality-control results, is disclosed in Appendix E. This revision adds no new model experiments, human adjudication, or benchmark repair. Human authors remain responsible for the claims, source accuracy, release permissions, and final venue-specific disclosures.

Reproducibility Statement

Section 3 and Algorithm 1 specify the method; Section 4 and Appendix B specify the frozen execution and evaluation contracts. Appendix A states the formal assumptions and proofs. Appendix D maps quantitative claims to their aggregate artifacts, source identities, and reproduction commands. Public aggregates support the reported accounting, but regeneration of private per-question analyses requires authorized access to the frozen source reports.

References

Ananto Nayan Bala. Recall Is Not Enough: A Reader-Context Diagnostic for Budget-Constrained Retrieval-Augmented Generation. arXiv:2607.00725v2, 2026. URL https://arxiv.org/abs/2607.00725v2.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Borschinger, and Tal Schuster. Tomayto, Tomahto. Beyond Token-level Answer Equivalence for Question Answering Evaluation. In *Proceedings of EMNLP*, pp. 291–305, 2022. doi: 10.18653/v1/2022.emnlp-main.20. URL https://aclanthology.org/2022.emnlp-main.20/.

Bernal Jimenez Gutierrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. arXiv:2502.14802v2, 2025. URL https://arxiv.org/abs/2502.14802v2.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *Proceedings of COLING*, pp. 6609–6625, 2020. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Yi Jiang, Sendong Zhao, Jianbo Li, Bairui Hu, Yanrui Du, Haochun Wang, and Bing Qin. OptiSet: Unified Optimizing Set Selection and Ranking for Retrieval-Augmented Generation. arXiv:2601.05027v1, 2026. URL https://arxiv.org/abs/2601.05027v1.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O. Arik. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG. arXiv:2410.05983v1, 2024. URL https://arxiv.org/abs/2410.05983v1.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient Context: A New Lens on Retrieval Augmented Generation Systems. arXiv:2411.06037v3, 2024. URL https://arxiv.org/abs/2411.06037v3.

Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. Evaluating Open-Domain Question Answering in the Era of Large Language Models. In *Proceedings of ACL*, pp. 5591–5606, 2023. doi: 10.18653/v1/2023.acl-long.307. URL https://aclanthology.org/2023.acl-long.307/.

Tomas Kocisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, and Edward Grefenstette. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://doi.org/10.1162/tacl_a_00023.

Tom Kwiatkowski et al. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7, 2019. doi: 10.1162/tacl_a_00276. URL https://doi.org/10.1162/tacl_a_00276.

Dahyun Lee, Yongrae Jo, Haeju Park, and Moontae Lee. Shifting from Ranking to Set Selection for Retrieval Augmented Generation. In *Proceedings of ACL*, pp. 17606–17619, 2025. doi: 10.18653/v1/2025.acl-long.861. URL https://aclanthology.org/2025.acl-long.861/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of ACL*, pp. 9802–9822, 2023. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Siyuan Meng, Junming Liu, Yirong Chen, Song Mao, Pinlong Cai, Guohang Yan, Botian Shi, and Ding Wang. From Ranking to Selection: A Simple but Efficient Dynamic Passage Selector for Retrieval Augmented Generation. arXiv:2508.09497v1, 2025. URL https://arxiv.org/abs/2508.09497v1.

Dongyu Ru et al. RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation. arXiv:2408.08067v2, 2024. URL https://arxiv.org/abs/2408.08067v2.

Alireza Salemi and Hamed Zamani. Evaluating Retrieval Quality in Retrieval-Augmented Generation. arXiv:2404.13781v1, 2024. URL https://arxiv.org/abs/2404.13781v1.

Hao Sun, Hengyi Cai, Yuchen Li, Xuanbo Fan, Xiaochi Wei, Shuaiqiang Wang, Yan Zhang, and Dawei Yin. Enhancing Retrieval-Augmented Generation via Evidence Tree Search. In *Proceedings of ACL*, pp. 24116–24127, 2025. doi: 10.18653/v1/2025.acl-long.1175. URL https://aclanthology.org/2025.acl-long.1175/.

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://doi.org/10.1162/tacl_a_00475.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving Retrieval-Augmented LMs with Compression and Selective Augmentation. arXiv:2310.04408, 2023. URL https://arxiv.org/abs/2310.04408.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of EMNLP*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Tao Yuan et al. LV-Eval: A Balanced Long-Context Benchmark with 5 Length Levels Up to 256K. arXiv:2402.05136, 2024. URL https://arxiv.org/abs/2402.05136.

A Formal Details

A.1 Witness availability and mixed marginal returns

Let the witness family be $\{\{a, b\}, \{c, d\}\}$. Adding b to $\varnothing$ changes availability by zero, while adding it to $\{a\}$ changes availability by one, violating submodularity. Adding d to $\{c\}$ changes availability by one, while adding it to $\{a, b, c\}$ changes availability by zero, violating supermodularity. Availability remains monotone.

These statements concern the exact objective in Equation (1), not every useful surrogate for evidence selection. They assume a fixed interpretation: adding a passage does not retract an already valid witness. A contradictory context can still contain a valid subset while becoming harder for a reader to use. Monotone witness availability therefore does not imply monotone reader utility.

A.2 Recall–completeness bound

Within a fixed-h stratum, let $X = |H_q \cap S_q|$. Then $r = \mathbb{E}[X]/h$ and $c = \Pr(X = h)$. Pointwise,

$$\mathbf{1}[X = h] \leq \frac{X}{h} \leq \frac{h-1}{h} + \frac{\mathbf{1}[X = h]}{h}.$$

Taking expectations gives Equation (2). Mixing full recovery and zero recovery attains the upper endpoint. When $r \geq (h-1)/h$, mixing recovery counts h and $h-1$ attains the lower endpoint. Below that threshold, counts at most $h-1$ can realize the desired recall with zero completeness. Additional budget constraints can restrict feasibility; if $k < h$, completeness is zero. For $h = 1$, both bounds coincide. The bound is an elementary measurement constraint, not an optimality result for CAST.

A.3 Opportunity monotonicity

If $C \subseteq C'$ and $k \leq k'$, every subset feasible under (C, k) remains feasible under (C', k'). Maximizing the same utility over the expanded feasible set proves monotonicity of V_R. The utility-difference identity follows from $U_R = V_R - G_R$.

The proof holds the utility definition and per-subset rendering fixed. Changing the reader contract changes the optimization problem. Neither the proof nor the identity certifies the implemented exactly-k selector or estimates how much answer utility is lost when official support is discarded.

B Implementation and Evaluation Contracts

B.1 Native retrieval and query generation

The native graph ordering is computed for each question, not stored as an offline relevance ranking. Qwen2.5-3B-Instruct provides entity/triple extraction, and NV-Embed-v2 provides passage, entity, and fact embeddings. Query-conditioned personalized PageRank and the bounded semantic walk produce native candidates, an incumbent final five, and frontier hypotheses with path evidence.

The generator receives bounded retrieved-text excerpts, not a separately evaluated abstractive summary. Its validation and bounded repair routines belong to the frozen implementation. Queries target unresolved relations and must satisfy both length limits: at most eight words and 96 characters. Contexts may repeat across frontier hypotheses; Equation (5) deduplicates scoring contributions while preserving distinct frontier/query provenance. Neither the generator nor the selector receives corpus answer labels.

B.2 Packing and explicit fallbacks

The proposed pool and effective selector pool are recorded separately. If a prompt exceeds ceiling L, packing first reduces generated-query provenance, then removes tail candidates while preserving the native ten-passage reader prefix (Reader@10), the incumbent final five, and up to five leading generated candidates. Passage text is never truncated. An input that cannot satisfy these constraints is rejected.

The selector must return only the declared JSON label object. Malformed output, duplicate or unknown labels, or an incorrect number of labels triggers the declared fallback: the native incumbent in headline evaluation, or the first k distinct pool passages in the controlled study. These outcomes remain in reported performance. The different fallbacks preserve the meaning of non-native and variable-k ablations.

In the fresh matrix, 41 of 8,000 joint prompts use the fallback. Five require reduced provenance, none prune passage candidates, and the largest effective selector input is 11,968 tokens under a 12,000-token ceiling. The native path and provenance are selector inputs, not additional reader passages. Generated label order determines neither passage order nor additional text for the reader.

B.3 Model, reader, and statistical contracts

The same Qwen3.6-27B snapshot performs selection and reading, using greedy decoding, 4-bit NF4 quantization, and bfloat16 computation. Selector thinking is disabled and its output ceiling is 128 tokens. Selection preserves pool order.

The reader's output ceiling is 1,024 tokens. It produces a rationale before an `Answer:` marker, and evaluation extracts the first line after the final marker. A missing or empty answer becomes `UNKNOWN`. MuSiQue uses a 12,000-token input ceiling; the other headline cohorts use 13,000 after an input-only check performed before their outcomes were produced. Conditions within each comparison share rendering, decoding, and extraction. Marker and generation-limit outcomes describe this implementation, not an ideal reader.

The controlled study has 12 fresh conditions, two historical reference rows, and one identity alias. Fresh comparisons use contemporaneous reader inference. Computation is reused only for the same question with identical ordered evidence. Pointwise scoring takes values from 0 to 9, breaks ties by pool position, and restores pool order.

The original ablation contrasts were declared before execution. Paired question-level percentile bootstrapping uses 20,000 replicates and seed 20260814. Pool-feasibility accounting, support-count strata, the budget interaction, and marker accounting are exploratory. Intervals are nominal and unadjusted; resampling questions does not establish transfer to new source documents, model families, or repeated-inference runs.

B.4 Resource accounting and sign tests

The fresh matrix uses 43,316 logical prompts and 12,578 physical generation calls. Reused upstream captures record 6,809 generation attempts and 24,624 retrieval events, but not token totals. Joint and pointwise selector output-token counts are 15,046 and 52,320. Corresponding reader input-token totals are 1,490,256 and 1,498,014. Input totals include rendered prompts, not just passage text; wall time is not a controlled latency comparison.

The selector input totals in Section 5.5 are 6,086,903 for joint selection and 9,156,239 for pointwise selection, over 1,000 and 26,160 logical prompts, respectively. Batching and sequence lengths differ. Missing upstream token totals preclude an end-to-end compute claim at matched search budget.

Exact sign-test p-values for headline F1 directions are 3.0×10^{-16} (MuSiQue), 5.9×10^{-19} (2WikiMultiHopQA), 0.0029 (HotpotQA), 0.880 (NaturalQuestions), 0.289 (PopQA), 0.031 (NarrativeQA), and 0.064 (LV-Eval). These test the direction of discordant pairs, not the mean magnitude estimated by the bootstrap.

C Supplementary Controlled Results

C.1 Full opportunity partitions

Increasing native width from cap 10 to cap 20 with joint selection and a five-passage reader raises F1 by 3.17 points [1.54, 4.81]. Available-and-fitting sets increase by a net 98 and preserved sets by 44. Moving from native cap 20 to mixed cap 35 adds a net 143 opportunities and 86 complete selections, with the 3.17-point F1 gain [1.21, 5.14] in Table 1. The second contrast changes width

Table C1: Annotation-completeness partitions, with 1,000 questions in every row. Generated-only candidates still depend on graph-conditioned query generation. The cap-10 machine-readable arm is named `protected_joint_10_to_5` but contains native candidates.

Pool and selector	Absent	Over budget	Feasible but discarded	Preserved
Native cap 10, joint 5	387	0	38	575
Native cap 20, joint 5	289	0	92	619
Generated cap 35, joint 5	292	0	100	608
Unprotected cap 35, joint 5	148	0	158	694
Mixed cap 35, pointwise 5	146	0	332	522
Mixed cap 35, joint 5	146	0	149	705
Mixed cap 35, joint 3	146	74	225	555
Mixed cap 35, joint 8	146	0	92	762
Mixed cap 50, joint 5	143	0	150	707

Table C2: Remaining original contrasts. The generated-only arm achieves 60.00 F1; mixing increases support recall by 5.03 points and full support by 9.70 points. Because generation remains graph-conditioned, this is not a graph-free comparison. Protecting native availability is an operational contract, not an independently established gain mechanism.

Intervention	F1 difference [95% CI]	Interpretation
Generated-only cap 35 $\rightarrow$ mixed cap 35	$+0.49$ $[-1.46, +2.41]$	Incremental answer effect of native candidates unresolved
Unprotected $\rightarrow$ protected mixed pool	$+0.15$ $[-1.72, +2.01]$	No demonstrated aggregate F1 benefit from protection
Mixed cap 35 $\rightarrow$ cap 50	-0.22 $[-0.76, +0.29]$	Small increase in realized width and opportunity

and source composition together. These are accounting differences, not causal fractions of newly discovered evidence.

C.2 Remaining source and width contrasts

The cap-50 intervention raises mean width from 26.16 to 27.48 and available complete support from 854 to 857. Its null F1 interval and 987 F1 ties do not establish general retrieval saturation or imply that answers are identical. The pool cap is not the reader passage budget.

C.3 Completion accounting

Let m_P, m_J indicate answer-marker presence for pointwise and joint selection at $k = 5$, and let $F_{q,P}, F_{q,J}$ be normalized per-question F1. Each row contributes the sum of $(100/n)(F_{q,J} - F_{q,P})$ within its marker-transition group. The denominator is the full cohort of 1,000 questions, not the group size.

Table C3: Post-treatment accounting, not causal mediation. Unrounded contributions sum exactly; separately rounded values can differ by 0.01. Marker-discordant groups contribute 4.57 F1 points, or 75.03% of the total, using unrounded values.

Pointwise marker	Joint marker	Questions	Contribution to total F1 difference
Absent	Absent	135	0.00
Absent	Present	111	$+7.55$
Present	Absent	57	-2.97
Present	Present	697	$+1.52$
Total		1,000	$+6.10$

Table C4: Percentages for the shared mixed pool and five-passage output. Strata count annotated support documents, not independently labeled reasoning hops.

Annotated support count	n	Pointwise recall	Pointwise complete	Joint recall	Joint complete
2	518	88.61	77.99	94.40	89.58
3	316	74.16	35.76	87.03	66.46
4	166	49.55	3.01	62.95	18.67

The 507-question both-complete subset has joint-minus-pointwise F1 of -1.29. It is another post-treatment partition, not a fixed-evidence comparison. These analyses do not apportion the total benefit among bridge preservation, distractor reduction, calibration, and completion.

C.4 Support-count strata

Joint selection improves completeness in every stratum, but the improvement is not monotone in support count; it is largest for three-support questions. On four-support questions, 62.95% joint recall coexists with 18.67% complete recovery. Equation (2) permits a broad completeness range at that recall. Reporting both quantities exposes difficulty that average recall hides, without isolating complementarity from general question difficulty as the cause of an answer gain.

C.5 Paired seven-cohort effects

Table C5: Within-reader paired effects for our frozen whole-pipeline runs. CAST's mean F1 exceeds our HippoRAG 2-derived reproduction on all seven packaged cohorts; six paired intervals exclude zero. These are complete packaged cohorts, not full upstream splits. The fresh MuSiQue controls in Table 1 are separate reader runs.

Dataset	n	Native F1	CAST F1	Difference [95% CI]
MuSiQue	1,000	48.70	60.05	$+11.35$ $[+8.80, +13.91]$
2WikiMultiHopQA	1,000	73.20	80.78	$+7.58$ $[+5.90, +9.33]$
HotpotQA	1,000	76.99	79.49	$+2.50$ $[+1.33, +3.73]$
NaturalQuestions	1,000	61.29	63.47	$+2.18$ $[+0.57, +3.83]$
PopQA	1,000	60.73	62.14	$+1.41$ $[+0.11, +2.72]$
NarrativeQA	293	26.51	30.24	$+3.73$ $[+0.82, +6.67]$
LV-Eval	124	12.53	15.57	$+3.04$ $[-1.93, +8.02]$

The table preserves the within-reader comparisons that underlie the aggregate transfer claim. It is separate from the cross-reader literature context in Table 4.

C.6 Reader-budget endpoints

Table C6: Detailed reader-budget endpoints for the ablation study in Table 3. More passages improve native-ranking F1, while the CAST five-to-eight contrast remains unresolved. Joint selections are requested anew at each k; they are not nested additions to one fixed context.

Evidence pipeline	Reader k	F1	Full support	Lexical coverage
Native ranking	5	48.59	43.90	68.20
Native ranking	8	51.63	57.40	78.30
Native ranking	10	54.80	61.30	82.00
Mixed pool, joint selection	3	59.75	55.50	79.60
Mixed pool, joint selection	5	60.48	70.50	86.00
Mixed pool, joint selection	8	60.07	76.20	89.20

The main-paper ablation table reports paired changes. These endpoint values show that additional passages increase annotated support and lexical coverage even when the corresponding CAST F1 contrast remains unresolved.

C.7 Why F1 divided by coverage is not reader accuracy

For F1 F and lexical-coverage indicator L,

$$\mathbb{E}[F] = \Pr(L = 1)\mathbb{E}[F \mid L = 1] + \Pr(L = 0)\mathbb{E}[F \mid L = 0].$$

Dividing by $\Pr(L = 1)$ leaves the second term unless uncovered questions contribute zero F1. Even then, lexical coverage does not establish grounding. Legacy ratio artifacts remain for provenance, not as evidence for a reader-conversion law or a semantic ceiling.

D Reproducibility and Claim-to-Artifact Map

Claim or procedure	Committed artifact
Paired controlled gains, intervals, and supplementary contrasts	`musique_paper_ablation_comparisons.csv`
Arm budgets, widths, metrics, and logical resources	`musique_paper_ablation_arms.csv`
Feasibility, support strata, budget interaction, and completion accounting	`paper_v4_analyses.json`
Frozen-output derivation	`analyze_paper_v4.py`
Whole-cohort results and sign tests	`headline_results.csv`
Audit counts, categorical labels, and quality control	`histograms`, `cases`, `QC`
Audit sample and annotation contract	`manifest`, `rubric`
Physical versus logical costs	`musique_paper_ablation_costs.csv`, `summary`
Frontier-conditioned query generation	`global_query_generator.py`
Duplicate-aware fusion, provenance, and packing	`generated_query_late_fusion.py`
Label validation and reader-order preservation	`joint_evidence_compressor.py`
Complete method and ablation protocols	`METHOD.md`, `MUSIQUE_ABLATION.md`
Frozen models, corpus/source identities, and environment	`PROVENANCE.md`, `SOURCE_MANIFEST.json`, `REPRODUCING.md`

The private MuSiQue report SHA-256 is `02c3e781e51849877336d17987bc689d5c9877c9408a0f38db478208203bdf4f`. The Qwen3.6-27B snapshot is `6a9e13bd6fc8f0983b9b99948120bc37f49c13e9`. The exploratory derivation verifies the report hash before parsing, reconstructs support from document identities, and checks public arm marginals. Its release contains aggregates only.

Table E1 uses final per-dataset failure-family counts. Its last column divides `scoring_equivalence` + `benchmark_ambiguity` by all non-success cases. Cross-dataset ratios sum counts before division. Neither `no_fault_ceiling` nor `residual_headroom` is interpreted as recoverable F1.

Authorized report holders can regenerate the exploratory aggregate from the source checkout:

```
PYTHONPATH=src python scripts/analyze_paper_v4.py \
  --report <authorized-frozen-MuSiQue-report> --check

python scripts/validate_repo.py
```

The first command recomputes the analysis against its immutable input. The public validator checks pinned aggregate and derivation identities; it does not recover private per-question intervals. All Version 9 results come from the same frozen experiments and analyses as Version 7. The revised organization, terminology, and explanatory structure introduce no new experimental evidence, human adjudication, or completed benchmark repair.

E Diagnosing the Errors That Remain

E.1 Locked audit protocol

The audit extends the found–preserved–used distinction to residual errors. It samples 100 questions without replacement from each packaged cohort using seed 20260824. Frozen traces supply objective diagnostics; a fixed rubric assigns one primary failure family to each non-full-F1 output.

Two separate `gpt-5.6-sol` contexts independently label 140 cases. Agreement is 90% for both primary stage and failure family, with kappas of 0.850 and 0.863, respectively. Within those 140 double-annotated cases, 108 have a disagreement on at least one annotation axis and are adjudicated by `claude-opus-5`. Agreement on stage/family and agreement on every axis are different quantities.

Full-F1 outputs are called successful by rule, not checked for false acceptance. Equal-sized dataset strata do not estimate prevalence weighted by dataset size or application traffic. The labels are model-assisted diagnoses, not independently established human ground truth.

Table E1: The four count columns partition 100 audited outputs per dataset. The final column is a conditional share of model-assigned labels, not a corrected score, saturation percentage, or performance ceiling.

Dataset	Full F1	Scoring equivalence	Reference ambiguity	System-actionable	Scoring/reference share of non-full F1
MuSiQue	47	19	20	14	39/53 (73.6%)
2WikiMultiHopQA	76	14	6	4	20/24 (83.3%)
HotpotQA	65	26	8	1	34/35 (97.1%)
NaturalQuestions	48	17	27	8	44/52 (84.6%)
PopQA	56	30	5	9	35/44 (79.5%)
NarrativeQA	10	38	9	43	47/90 (52.2%)
LV-Eval	9	6	18	67	24/91 (26.4%)

Of 389 non-full-F1 cases, 243 receive scoring/reference labels and 146 receive system-actionable labels. The latter include 61 selection failures, 51 reader reasoning/completion failures, and 34 combined frontier, generated-retrieval, or admission failures. Among the five support-labeled cohorts, 174 of 208 non-full-F1 cases retain lexical answer coverage and 116 retain all official support passages. These observations must not be collapsed into one notion of success.

E.2 Metric saturation is not task saturation

On the full packaged 2WikiMultiHopQA and HotpotQA cohorts, complete support is 97.70% and 98.00%, while answer F1 is 80.78 and 79.49. This leaves little numerical room on that annotated-support metric for this system. It does not establish that answers are always justified, that readers cannot improve, or that the intended tasks are solved.

A second issue is measurement. Scoring/reference labels account for 172/208 (82.7%) non-full-F1 cases in the five support-labeled audit cohorts. Token matching can reject acceptable alternatives, while semantic judges can accept unsupported additions (Bulian et al., 2022; Kamalloo et al., 2023). The audit therefore motivates validation of measurement, not replacement of a lexical score by an assumed infallible judge.

Genuine task saturation would require evidence that little independently validated capability remains to improve on the intended distribution. These experiments establish no such ceiling. One actionable HotpotQA case in a 100-case sample does not imply that only one percent of the benchmark is improvable. Partial credit, sampling error, competing causes, and unaudited full-score outputs prevent that interpretation.

E.3 Dataset-specific research priorities

MuSiQue. Assembly remains difficult. Full support is 70.50% overall and 18.67% for four-support questions. Future studies should retain support-count strata and distinguish candidate absence from feasible selection loss.

2WikiMultiHopQA and HotpotQA. Near-complete annotated support makes these cohorts useful preservation checks. Harder capability claims need renewed evidence: checked references, grounding judgments, and fresh challenge items rather than automatically treating every residual answer error as a retrieval failure.

NaturalQuestions and PopQA. Lexical coverage is 95.50% and 90.00%, but full annotated support is 49.30% and 42.20%. Answer presence and justified use remain separate. PopQA also has five generated-retrieval misses among nine audited actionable cases; high answer-string coverage does not eliminate retrieval headroom.

NarrativeQA and LV-Eval. Together these cohorts contribute 110 of 146 actionable cases and 56 of 61 selection failures (15 and 41, respectively). Evidence-selection failures concentrate in this long-context regime. NarrativeQA also has 47 scoring/reference labels, so system and measurement improvements are complementary. Alternative-support annotations would be especially useful because these packaged evaluations lack official support labels.

These priorities are conditional on our system, rubric, and cohorts, not upstream splits or stronger readers. The audit covers 100 of 124 LV-Eval cases and 100 of 293 NarrativeQA cases. The exposed samples should not serve as untouched confirmation sets for methods developed from their failures.

E.4 Benchmark renewal without breaking compatibility

The earlier draft's proposal for independently validated V2.0 releases of all seven benchmarks remains future work. It follows from the distinct failure regimes, not a claim that this audit has already repaired the datasets. A common maintenance contract would include:

1. Blind human review of several systems' outputs, including rejected and full-score answers, to measure false rejection and false acceptance.
2. Versioned item/reference manifests distinguishing retained, extended, repaired, and retired examples; difficulty alone must not justify removal.
3. Pinned original and equivalence-aware scorers, with compatibility and system-ranking analyses rather than a silently more permissive metric.
4. Alternative-witness annotations and fresh, independently checked challenge sets, released separately from historical compatibility sets.

New prompts, thresholds, and training examples should be developed on separate, train-derived failures. Existing cohorts remain useful regression tests. Stronger capability claims should not rest only on small score differences whose grounding and reference validity are uncertain.