EGAL-SC: EVIDENCE-GAP-AWARE LOGICAL RETRIEVAL WITH SAFE CONTROL FOR MULTIMODAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval controls can recover missing evidence while excluding support needed to answer the same question. Effective retrieval therefore requires deciding which candidates should enter the answering context. We introduce EGAL-SC, a multimodal retrieval framework that pairs controlled and unrestricted searches with a state-conditioned evidence admission rule. The rule combines relevance, alignment with accumulated evidence, novelty, and a learned document-error signal. On 562 held-out first-control states from MultiModalQA and WebQA, threshold-free ranking of the same candidate pools under a common five-document cap yields 272 complete-evidence outcomes, compared with 239 for maximal marginal relevance, and retains 49 additional support units. The complete-evidence difference passes a four-test Holm correction; the smaller difference from removing the error signal does not. In a seven-system comparison on two 500-question panels, EGAL-SC obtains 21.933% all-question F1 versus 19.520% for Agentic Hybrid, with broader answer coverage and higher computation. On jointly answered questions, its F1 is lower. These results support state-conditioned evidence ranking and show that system-level answer gains must be assessed together with answer selection and cost.

1 INTRODUCTION

Finding a new source is not the same as improving the evidence available for an answer. A constrained search may expose a missing image while excluding a text passage needed to interpret it; a rewritten query may retrieve new facts but displace an earlier supporting source. For a retrieval agent, the useful decision is therefore not only *where to search*, but also *which results to retain in the answering context*.

Retrieval-augmented generation connects language models to external knowledge (Lewis et al., 2020; Guu et al., 2020; Izacard & Grave, 2021). Logical search and direct corpus interaction make that access more expressive (Zeng et al., 2026; Li et al., 2026; Subramanian et al., 2025). Yet a valid search action need not produce a useful replacement for the current evidence. Query relevance alone overlooks the relation between a candidate and the support the agent already possesses; diversity alone does not establish that a document is factually reliable.

EGAL-SC separates proposing a retrieval control from admitting its results. Matched unrestricted and controlled requests expose alternatives for the same retrieval intent. A state-conditioned rule then combines candidate relevance, alignment to the accumulated evidence, novelty relative to individual admitted sources, and a learned document-error signal. The reference state is fixed during each admission decision. This makes candidate acceptance an explicit operation rather than an automatic consequence of executing a search.

We evaluate evidence selection and complete-agent behavior. A held-out comparison on 562 first-control states tests evidence ranking against MMR and a no-risk variant using common candidates and a five-document cap. A seven-system comparison on two 500-question panels examines answer coverage, answer quality, and computation. The state comparison changes the selector within a fixed context; the system comparison evaluates complete configurations.

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.

The contribution is a state-conditioned admission framework evaluated through support preservation and downstream answering. Joint measurement of support identities and complete-evidence outcomes makes both recovery and regressions visible. The strongest ranking result is the complete-evidence difference from MMR; the incremental risk effect is smaller. The system comparison further shows why aggregate answer gains require care: F1 increases overall while decreasing on jointly answered questions.

2 Related Work

Retrieval and control. Dense retrieval, late interaction, and sparse expansion improve candidate access (Karpukhin et al., 2020; Khattab & Zaharia, 2020; Santhanam et al., 2022; Formal et al., 2021); their effectiveness varies across tasks (Thakur et al., 2021). BM25, rank fusion, and hypothetical-document retrieval supply complementary rankings (Robertson & Zaragoza, 2009; Cormack et al., 2009; Gao et al., 2023). LOGICALRAG, direct corpus interaction, and agentic keyword search expose richer search interfaces (Zeng et al., 2026; Li et al., 2026; Subramanian et al., 2025). EGAL-SC instead examines the admission decision after a control has produced candidates.

Relevance and diversity. Maximal Marginal Relevance (MMR) balances query relevance against redundancy with already selected items (Carbonell & Goldstein, 1998). DF-RAG adapts diversity to the query by evaluating candidate evidence sets (Khan et al., 2026). EGAL-SC shares the relevance–novelty motivation, but its alignment and novelty refer to accumulated cross-turn evidence, fixed within each admission call. A learned document-error score provides additional supervision. The relevant comparison is therefore whether these signals improve support preservation beyond ordinary diversification, not whether diversity itself is novel.

Reasoning and memory. ReAct, interleaved retrieval–reasoning, Self-Ask, active retrieval, and Self-RAG connect search with intermediate reasoning and self-assessment (Yao et al., 2022; Trivedi et al., 2023; Press et al., 2023; Jiang et al., 2023; Asai et al., 2023). RAPTOR, GraphRAG, and HippoRAG organize memory through hierarchies or relations (Sarthi et al., 2024; Edge et al., 2024; Jiménez Gutiérrez et al., 2024). Our focus is the selection boundary between a proposed search and the answering context, rather than a new memory representation.

Multimodal grounding. CLIP, Flamingo, BLIP-2, and visual instruction tuning connect image and language processing (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023). MuRAG and REVEAL combine multimodal retrieval with generation (Chen et al., 2022; Hu et al., 2022); MultiModalQA and WebQA support multimodal question answering (Talmor et al., 2021; Chang et al., 2021). Parametric errors and irrelevant context complicate the relation between retrieval and answers (Mallen et al., 2023; Yoran et al., 2023; Wang et al., 2024). We accordingly distinguish visible support, answer coverage, and quality on common answered questions.

3 Problem Setting

Let q_i be a question and $\mathcal{D}$ a corpus of text, tables, images, and derived fields. At retrieval cycle t, the agent observes history H_t, admitted-document memory E_t, and textual summaries M_t. A control changes access to candidates; admission selects documents for presentation; the answering policy produces a response or abstains. Reference answers and support labels are available only to the offline evaluator.

Let S_i be the question-indexed reference support set. A unit is a factual span, table row, or image-associated fact. For presented evidence E, $U(E)$ contains the reference units whose content is actually visible. Several units may belong to one document; selecting its identifier does not recover content omitted by presentation. Complete presented evidence is

$$c_i = \mathbf{1}\{S_i \neq \varnothing \ \wedge \ S_i \subseteq U(E_i^{\mathrm{fin}})\}. \tag{1}$$

This is an evidence measure, not answer correctness. In the name *Safe Control*, safety refers to guarded admission evaluated through support loss. At the evaluated setting $K = m = 5$, the entire five-document list can change; there is no protected nonzero prefix of the plain list.

4 METHOD

EGAL-SC follows three principles: retain an unrestricted retrieval reference, select candidates in relation to the current evidence state, and separate admission from the decision to answer. Figure 1 summarizes their interaction.

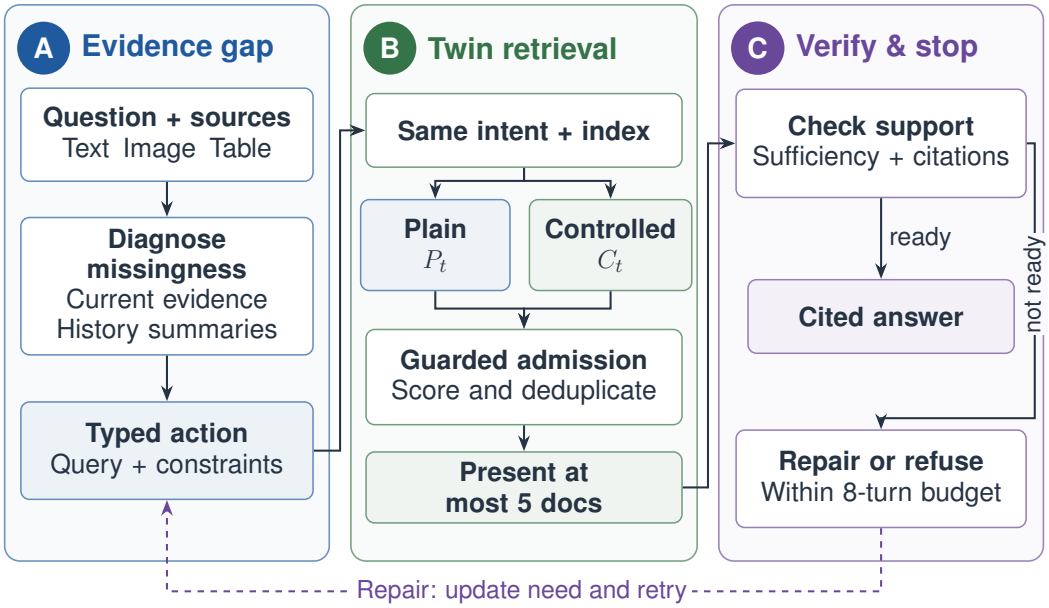

Figure 1: EGAL-SC separates search control from candidate admission. The diagram depicts normal verification and repair. Budget exhaustion and bounded rewrites can instead return an existing draft (Section 4.3); reference support labels remain offline.

4.1 MATCHED RETRIEVAL AND EVIDENCE STATE

The index exposes titles and text, OCR, captions, visual tags, table rows, and typed metadata, excluding answers, reference support IDs, and error labels. BM25 supplies lexical ranking; hybrid requests combine text and CLIP-based image rankings and retain at most 50 candidates. The action schema contains `action_type`, `query_expansion`, and `filter_condition`. Actions are `retrieve`, `filter`, or `refine`, using keyword, time, modality, and Boolean conditions. A deterministic compiler accepts supported fields; empty or invalid controls use the plain path.

For expanded intent q_t and control u_t, matched requests are

$$P_t = R(q_t), \qquad C_t = R(\mathcal{C}(q_t, u_t)) \,. \tag{2}$$

They share the index, ranking configuration, and candidate budget. Fusion occurs within each request; admission then scores the occurrence multiset $\mathcal{Q}_t = P_t \uplus C_t$ before deduplication. The memory E_t accumulates admitted document vectors with cross-turn ID deduplication. Answer-context summaries are stored separately. Both memories are held fixed in the held-out state comparison.

4.2 STATE-CONDITIONED CANDIDATE ADMISSION

Features and supervision. Each query or document has a text block $u(x) \in \mathbb{R}^{768}$ and visual block $v(x) \in \mathbb{R}^{768}$. The local encoders return final-layer CLS text features and mean image-patch features excluding the class token. Present blocks are individually L2-normalized; absent modalities are zero blocks. Concatenation gives $z_x = [u(x); v(x)] \in \mathbb{R}^{1536}$ without another normalization. The base score is $b_t(d) = \cos(z_{q_t}, z_d) \in [-1, 1]$; text-only queries have $v(q_t) = 0$. Appendix B.1 specifies the extraction interface.

The document-error head takes $x_{td} = [u(q_t); z_d] \in \mathbb{R}^{2304}$ and has affine widths $2304 \to 256 \to 128 \to 1$, with ReLU and dropout after hidden layers. The 12,000-pair annotation pool contains 9,600 training and 2,400 validation pairs. Labels identify erroneous or misleading document content under human review, not errors in generated answers. A negative label means no error was established. With encoders frozen, the head is trained using binary cross-entropy on logits and selected by validation loss. Its inference score is $r_t(d) = \sigma(h_\phi(x_{td}))$; training details appear in Appendix B.2.

Alignment, novelty, and eligibility. For a nonempty evidence memory with finite nonzero vectors and a nonzero mean, define

$$\bar{z}_t = |E_t|^{-1} \sum_{j \in E_t} z_j, \qquad g_t(d) = 1 - \cos(\bar{z}_t, z_d),$$
$$\widetilde{g}_t(d) = \frac{g_t(d) - g_t^{\min}}{g_t^{\max} - g_t^{\min}}, \qquad n_t(d) = 1 - \max_{j \in E_t} \cos(z_d, z_j). \tag{3}$$

The extrema use the valid candidate pool before thresholding. Raw gap and novelty lie in $[0, 2]$; only gap is min–max normalized. The score is

$$s_t(d) = 0.20 b_t(d) + 0.30(1 - \widetilde{g}_t(d)) + 0.25 n_t(d) + 0.25(1 - r_t(d)). \tag{4}$$

Relevance links candidates to the current intent. Alignment uses the evidence mean, whereas novelty uses the closest admitted document: proximity to the aggregate and duplication of an individual source are different geometric properties. The error head supplies a supervised document-content signal. All three depend on available inputs, not missing Gold support. The memory E_t stays fixed throughout an admission call, so scores do not change after each selected document.

Candidates are scored, deduplicated by ID, filtered at the fixed threshold $s_t \geq 0.65$, and truncated to at most five. Ties use decreasing base score and then increasing ID. Rejected positions are not refilled. Algorithm 1 also defines initialization and degenerate cases. An unavailable risk estimate receives $r = 1$, removing its positive complement without guaranteeing rejection. Appendix B derives the risk eligibility boundary and gives the full validity rules.

Algorithm 1 Initialization and candidate admission

Require: Ordered P_t, C_t; control status; memory E_t; features; $K = 5$, $\tau = 0.65$.
Ensure: At most K ordered document identifiers.
1: On first retrieval, **return** top K distinct valid plain candidates by $(-b_t, \mathrm{id})$.
2: **if** control is empty or invalid **then**
3: **return** first K distinct valid identifiers of P_t, preserving retrieval order.
4: **end if**
5: Form $\mathcal{Q}_t = P_t \uplus C_t$; set unavailable risk to 1. Exclude parse failures, zero candidate vectors, or non-finite base scores; if empty, **return** $[\,]$.
6: **if** $E_t = \varnothing$ or $\bar{z}_t = 0$ **then**
7: **return** top K distinct valid candidates by $(-b_t, \mathrm{id})$.
8: **end if**
9: Compute g_t on the valid pool; obtain $g_t^{\min}, g_t^{\max}$.
10: **if** $g_t^{\min} = g_t^{\max}$ **then**
11: **return** top K distinct candidates by $(-b_t, \mathrm{id})$.
12: **end if**
13: Compute n_t, r_t, s_t; reject non-finite scores; refresh extrema if the pool changes.
14: Deduplicate by greatest s_t, then greatest b_t; retain $s_t \geq \tau$.
15: Sort by $(-s_t, -b_t, \mathrm{id})$; **return** first K; do not refill rejected positions.

Ranking diagnostic. Full–Rank uses Eq. (4) and the same validity, deduplication, and tie rules, but omits the $s_t \geq 0.65$ filter. The no-risk variant removes the entire $0.25(1 - r_t)$ contribution and leaves all other relative coefficients unchanged. Both rank up to five documents. This diagnostic compares the ordering supplied by the signals; it does not treat a threshold-free output as evidence that the threshold itself is beneficial.

4.3 PRESENTATION AND THE DECISION TO ANSWER

The answerer sees selected sources and delivered historical summaries. Only visible content contributes to measured support. An instruction-guided LLM evaluates the question, evidence summaries, and a draft, returning a self-assessed score $a \in [0, 100]$; online sufficiency is $\widehat{c} = a/100$. Reference labels are excluded. The threshold 0.65 acts on this self-report, not on measured reference coverage.

The loop monitors new strings by splitting admitted text at full stops and semicolons and exact-matching spans against its inventory. No new span increments a stagnation counter; a new span resets it. Initial plain retrieval and base-ranked selection count toward the eight-cycle limit. Empty admission preserves existing evidence and permits repair while budget remains.

Termination checks budget exhaustion first, two-cycle stagnation second, and citation/rewriting next. Budget exhaustion returns an eligible draft; stagnation otherwise yields refusal. Normal early answering requires $\widehat{c} \geq 0.65$ and passing source-reference checks. Citation failure permits two rewrites with reassessment, after which draft fallback is possible. Eligible drafts are nonempty, parseable, and non-abstaining, even without usable sufficiency or passing citations. An empty eligible set yields refusal. Normal verification and draft fallback are distinct exits. Appendix B.4 gives a deterministic fallback selector: scored drafts enter a highest-score band before citation-based tie-breaking, while entirely unscored sets use retrieval relevance.

5 EVALUATION

5.1 EVALUATION POPULATIONS AND COMPARISONS

Evidence selection. A panel of 600 held-out questions contains 300 MultiModalQA and 300 WebQA questions. The controller reaches usable first-control states on 283 and 279 questions, respectively. These 562 states are the evaluation population; 38 questions do not reach a usable state. The held-out questions do not overlap the training, development, or system-evaluation questions. At each state, history, candidate features, retrieval results, and rendering are common across selectors. All selectors use a five-document cap. The controller is unchanged apart from selector interfaces, so the comparison measures new-question generalization, not transfer to a different language model.

MMR greedily trades current-query similarity against redundancy with documents already selected in the same call (Carbonell & Goldstein, 1998). Its first item has the highest base similarity; subsequent ties use base score and then document ID. The development procedure selects $\lambda = 0.6$ on 50 disjoint states, prioritizing complete evidence, fewer loss states, retained support, and finally the smaller parameter. Full–Rank and its no-risk variant retain the original relative weights. No parameters are retuned on the held-out questions (Appendix C).

Complete agents. Two 500-question panels compare seven complete system configurations, including EGAL-SC and Agentic Hybrid. The controller, answerer, verifier, and offline judge are distinct roles using a private Qwen2.5-family instruction backbone. Retrieval interfaces, feature settings, and serving conditions are specified in Appendix A. This comparison measures system-level behavior; it does not isolate admission from retrieval, verification, or stopping.

5.2 OUTCOMES AND STATISTICAL ANALYSIS

For a fixed state let $R_i^0 = S_i \cap U(\mathcal{R}(P_i[:5]))$ and $V_i^\pi = S_i \cap U(\mathcal{R}(A_i^\pi))$, where $\mathcal{R}$ is the common renderer. Lost and newly recovered support are

$$L_i^\pi = |R_i^0 \setminus V_i^\pi|, \qquad G_i^\pi = |V_i^\pi \setminus R_i^0|. \tag{5}$$

The count identity is $|V_i^\pi| = |R_i^0| - L_i^\pi + G_i^\pi$. A loss outcome is $L_i^\pi > 0$: replacing an old unit with a different new one can preserve the number of units while losing an initial support identity. We report loss incidence, complete-evidence outcomes from Eq. (1), and retained-unit totals.

Answer F1 uses lowercased, stopword-filtered word tokens, multiset overlap, and the best score over reference answers; abstentions score zero. It is the same study scorer on both datasets, not the official WebQA leaderboard scorer. For panel d, write $F_d = \sum_i f_i$, with $f_i \in [0, 1]$. Across the two

equal 500-question panels, $\mathrm{F1}_{\mathrm{all}} = 100(F_{\mathrm{MMQA}} + F_{\mathrm{WebQA}})/1000$ and coverage is $100n_a/1000$. Consequently, $\mathrm{F1}_{\mathrm{all}} = (n_a/1000)\mathrm{F1}_{\mathrm{answered}}$.

We use exact paired tests (McNemar, 1947) and a four-test Holm family (Holm, 1979) for held-out CE/loss. For jointly answered questions in the system comparison, an F1 interval uses paired difference moments within dataset strata and a t approximation. These retrospective analyses condition on independent question pairs. Appendices C.1 and E specify the calculations.

6 RESULTS

6.1 EVIDENCE SELECTION: COMPETITIVE GAINS WITH INDIVIDUAL REVERSALS

Full–Rank yields complete evidence on 272 of 562 states (48.40%), compared with 239 for MMR (42.53%), while retaining 437 rather than 388 support units (Table 1). The CE advantage is 16 states in MultiModalQA and 17 in WebQA. Retained support increases by 23 and 26 units, respectively. The improvement thus concerns both complete questions and retained evidence, rather than only the number of retrieved candidates.

Table 1: Held-out ranking on 562 first-control states. CE and loss count questions; retained support counts visible units. No parameters are retuned on these questions.

Selector	CE ↑	Loss ↓	Retained units ↑
MMR ($\lambda = 0.6$)	239	189	388
Full–Rank without risk	261	177	423
Full–Rank	272	171	437

The paired outcomes are not uniformly favorable: Full–Rank uniquely completes 57 states and MMR 24; support loss occurs only under Full–Rank in 27 states and only under MMR in 45. Under the independent-question model, CE passes the four-test Holm correction ($p = 0.001269$), whereas loss does not ($p = 0.088742$; Table 2). Thus the observed completeness gain has stronger statistical support than a reduction in loss incidence. The paired counts expose both benefits and regressions, rather than a uniform advantage across questions.

Table 2: Held-out paired differences, in percentage points, Full–Rank minus comparator. Exact McNemar tests and the complete four-test Holm family assume independent question pairs; they are not source-adjusted.

Comparator	Outcome	Difference	Exact p	Holm p
MMR	CE	$+5.872$	0.000317	0.001269
MMR	Loss	-3.203	0.044371	0.088742
No risk	CE	$+1.957$	0.019211	0.057632
No risk	Loss	-1.068	0.429591	0.429591

Removing the document-error signal gives 261 complete states, 177 loss states, and 423 retained units. Full–Rank adds 11 completions and 14 retained units, with six fewer loss states. Fifteen states are complete only with the signal, while four are complete only without it. The CE contrast has unadjusted $p = 0.0192$ but Holm-adjusted $p = 0.0576$. This smaller positive effect does not establish that the signal is necessary, nor does the lack of a corrected rejection establish equivalence. These comparisons evaluate threshold-free ordering, not the isolated effect of the 0.65 gate.

6.2 COMPLETE AGENTS: ANSWER QUALITY, COVERAGE, AND COMPUTATION

The seven-system comparison supplies a complementary diagnostic (Figure 2; Appendix D). EGAL-SC and Agentic Hybrid obtain 21.933% and 19.520% all-question F1. They jointly answer 307 questions, on which EGAL-SC scores 55.863% versus 58.697%. EGAL alone answers 138 questions

and Hybrid alone 23. With J, E, H denoting these sets,

$$\Delta \mathrm{F1}_{\mathrm{all}} = \frac{100}{1000}\left[\sum_{i\in J}(f_i^E - f_i^H) + \sum_{i\in E} f_i^E - \sum_{i\in H} f_i^H\right] = -0.870 + 4.783 - 1.500 = 2.413. \quad (6)$$

The nonoverlapping sets contribute a net $+3.283$ points, exceeding the negative common-answer contribution. Mean latency and tokens increase by 32.89% and 39.00%. The gain is concentrated in questions answered by only one system, while performance declines on the common answered set.

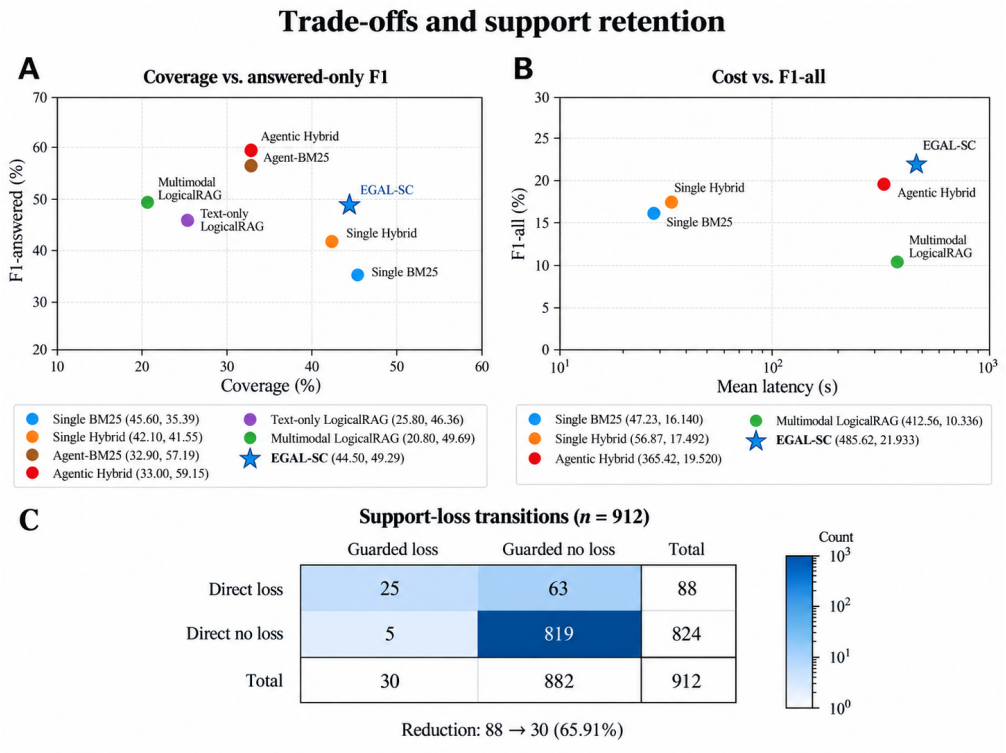

Figure 2: System-level answering and trajectory diagnostics. (A) Coverage versus answered-only F1. (B) Mean latency versus all-question F1, with logarithmic latency. (C) Endpoint loss transitions for Direct and guarded complete trajectories on 912 controlled questions. Endpoint loss measures the complete retrieval trajectory, including repair and verification.

7 Discussion and Conclusion

State-conditioned admission makes a useful distinction explicit: obtaining candidates, retaining support, and producing an answer are separate decisions. On held-out states, threshold-free ranking improves complete-evidence outcomes over MMR. Removing the document-error signal produces a smaller difference that does not pass the four-test Holm correction. At the system level, EGAL-SC increases all-question F1 and answer coverage, but uses more computation and scores lower on jointly answered questions.

The state distribution is controller-induced, and the support annotations and feature interfaces are laboratory maintained. The statistical calculations assume independent question pairs; dependence between questions sharing a source can change uncertainty. The ranking comparison does not isolate the admission threshold, and the system comparison does not identify the effect of an individual component. External execution requires access to the laboratory models and corpus.

The practical lesson is to measure what is retained, not merely what is retrieved, and to assess answer selection together with quality and cost. EGAL-SC improves context completeness in the evaluated states, while its system-level answer gains remain uneven across questions.

REPRODUCIBILITY STATEMENT

Section 4 defines admission; Appendix B specifies features, training, initialization, and fallback. Appendices C, D, and E provide aggregate outcomes, paired cells, and answer-set moments. The supplementary code reproduces the aggregate statistical calculations. Production execution depends on the laboratory's model and corpus access.

ETHICS STATEMENT

Human output assessments were blinded to system identity and disagreements were adjudicated separately. No assessor identities or restricted question content are released here. Retrieved sources can contain errors and bias. Source linkage, document-error scores, and fallback eligibility are not guarantees of factual correctness.

AI USE STATEMENT

ChatGPT and Codex assisted with code debugging, language editing, manuscript drafting and restructuring, methodological feedback, figure preparation, reference checks, arithmetic verification, record consistency checks, supplementary-code packaging, and interpretation. Experimental measurements were supplied by the laboratory; writing assistants did not execute the production benchmarks. The authors are responsible for the research decisions, experimental records, AI-assisted content, and final submission.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A Visual Language Model for Few-Shot Learning. *arXiv preprint arXiv:2204.14198*, 2022. doi: 10.48550/arXiv.2204.14198. Also published in NeurIPS 2022.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv preprint arXiv:2310.11511*, 2023. doi: 10.48550/arXiv.2310.11511.

Jaime Carbonell and Jade Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336. Association for Computing Machinery, 1998. doi: 10.1145/290941.291025.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. WebQA: Multihop and Multimodal QA. *arXiv preprint arXiv:2109.00590*, 2021. doi: 10.48550/arXiv.2109.00590. Revised 2022; CVPR camera-ready version.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In *Proceedings of EMNLP*, pp. 5558–5570. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.375.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of SIGIR*, pp. 758–759. Association for Computing Machinery, 2009. doi: 10.1145/1571941.1572114.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint arXiv:2404.16130*, 2024. doi: 10.48550/arXiv.2404.16130.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stephane Clinchant. SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval. *arXiv preprint arXiv:2109.10086*, 2021. doi: 10.48550/arXiv.2109.10086.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *Proceedings of ACL*, pp. 1762–1777. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.99.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-Augmented Language Model Pre-Training. *arXiv preprint arXiv:2002.08909*, 2020. doi: 10.48550/arXiv.2002.08909.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. doi: 10.2307/4615733.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A. Ross, and Alireza Fathi. REVEAL: Retrieval-Augmented Visual-Language Pre-Training with Multi-Source Multimodal Knowledge Memory. *arXiv preprint arXiv:2212.05221*, 2022. doi: 10.48550/arXiv.2212.05221. Revised 2023; also published in CVPR 2023.

Gautier Izacard and Edouard Grave. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of EACL*, pp. 874–880. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.74.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active Retrieval Augmented Generation. In *Proceedings of EMNLP*, pp. 7969–7992. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.495.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models. *arXiv preprint arXiv:2405.14831*, 2024. doi: 10.48550/arXiv.2405.14831. Also published in NeurIPS 2024.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of EMNLP*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550.

Saadat Hasan Khan, Spencer Hong, Jingyu Wu, Kevin Lybarger, Youbing Yin, Erin Babinsky, and Daben Liu. DF-RAG: Query-Aware Diversity for Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 2873–2894. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-eacl.150.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *arXiv preprint arXiv:2004.12832*, 2020. doi: 10.48550/arXiv.2004.12832. Also published in SIGIR 2020.

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. *arXiv preprint arXiv:2005.11401*, 2020. doi: 10.48550/arXiv.2005.11401. Also published in NeurIPS 2020.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*, 2023. doi: 10.48550/arXiv.2301.12597.

Zhuofeng Li, Haoxiang Zhang, Cong Wei, Pan Lu, Ping Nie, Yi Lu, Yuyang Bai, Shangbin Feng, Hangxiao Zhu, Ming Zhong, Yuyu Zhang, Jianwen Xie, Yejin Choi, James Zou, Jiawei Han, Wenhu Chen, Jimmy Lin, Dongfu Jiang, and Yu Zhang. Beyond Semantic Similarity: Rethinking Retrieval for Agentic Search via Direct Corpus Interaction. *arXiv preprint arXiv:2605.05242*, 2026. doi: 10.48550/arXiv.2605.05242.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485*, 2023. doi: 10.48550/arXiv.2304.08485. Also published in NeurIPS 2023.

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. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.546.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and Narrowing the Compositionality Gap in Language Models. In *Findings of EMNLP*, pp. 5687–5711. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.378.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*, 2021. doi: 10.48550/arXiv.2103.00020.

Stephen Robertson and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 4(1–2):1–174, 2009. doi: 10.1561/1500000019.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proceedings of NAACL-HLT*, pp. 3715–3734. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.272.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. *arXiv preprint arXiv:2401.18059*, 2024. doi: 10.48550/arXiv.2401.18059.

Shreyas Subramanian, Adewale Akinfaderin, Yanyan Zhang, Ishan Singh, Mani Khanuja, Sandeep Singh, and Maira Ladeira Tanke. Keyword Search Is All You Need: Achieving RAG-Level Performance without Vector Databases Using Agentic Tool Use. *arXiv preprint arXiv:2602.23368*, 2025. doi: 10.48550/arXiv.2602.23368.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. MultiModalQA: Complex Question Answering over Text, Tables and Images. *arXiv preprint arXiv:2104.06039*, 2021. doi: 10.48550/arXiv.2104.06039. Also published in ICLR 2021.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *arXiv preprint arXiv:2104.08663*, 2021. doi: 10.48550/arXiv.2104.08663. Also published in the NeurIPS 2021 Datasets and Benchmarks Track.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *Proceedings of ACL*, pp. 10014–10037. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.557.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, Ruicheng Yin, Changze Lv, Xiaoqing Zheng, and Xuanjing Huang. Searching for Best Practices in Retrieval-Augmented Generation. In *Proceedings of EMNLP*, pp. 17716–17736. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.981.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv preprint arXiv:2210.03629*, 2022. doi: 10.48550/arXiv.2210.03629. Revised 2023; ICLR camera-ready version.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. *arXiv preprint arXiv:2310.01558*, 2023. doi: 10.48550/arXiv.2310.01558.

Yuqi Zeng, Qixiang Deng, Yulei Wan, Ruiquan Jiang, Xiaoqing Zheng, and Xuanjing Huang. Rethinking Agentic RAG: Toward LLM-Driven Logical Retrieval Beyond Embeddings. *arXiv preprint arXiv:2605.27123*, 2026. doi: 10.48550/arXiv.2605.27123.

A Experimental Configuration and Evaluation Definitions

The seven-system evaluation uses a private Qwen2.5-family instruction backbone and a CLIP-based laboratory visual interface. Controller, answerer, verifier, and offline judge are distinct roles. Model labels below identify the evaluated laboratory configuration; they do not imply equivalence to a public checkpoint with the same family name.

Table 3: Retrieval interfaces and seven-system serving settings. Encoder labels refer to the laboratory feature interfaces defined in Appendix B.1.

Item	Setting
Lexical ranking	OpenSearch 3.7.1; BM25 $k_1 = 1.5$, $b = 0.75$
Lexical fields	Title weight 2.0; content weight 1.0
Within-request fusion	Text/image RRF: $0.6/(60+r_{\text{text}})+0.4/(60+r_{\text{image}})$; absent branches contribute zero
Candidate budget	At most 50 candidates per matched request after branch processing
Text feature input	Title plus first 512 body characters; one 768-dimensional CLS feature, then L2 normalization
Text encoder identifier	`bge-m3-v2.0-dense` (laboratory configuration label)
Image features	CLIP-ViT-L/14-labelled local interface, 336×336 input; 768-dimensional patch mean excluding the class token
Admission features	Text then visual block, 1536 dimensions; zero blocks for missing modalities; no second normalization
Base score	Expanded-query/document cosine of full 1536-dimensional vectors, without shifting or clipping
Risk signal	$2304 \to 256 \to 128 \to 1$ head, frozen encoders; sigmoid at inference; 9,600 training and 2,400 validation pairs
Admission	Weights 0.20/0.30/0.25/0.25; threshold 0.65; $K = m = 5$
Tie order	Decreasing admission score, decreasing base score, increasing document ID
Equal-distance branch	Base-score ranking instead of division by a zero gap range
Presentation	At most five selected documents; only delivered source and visible summary content counts toward support
OCR	PaddleOCR v2.8.0
Control	`retrieve/filter/refine`; keyword, time and modality fields; AND/OR/NOT
Stopping	Self-score threshold 0.65; two-round stagnation; eight retrieval cycles; at most two citation rewrites
Serving	Temperature 0, top-p 1; batch one; single concurrency; preloaded model
Hardware	One A800 80-GB GPU; two Xeon Gold 6330 CPUs; 512 GB memory

Presentation and support mapping. The support unit is a single question-relevant factual unit, represented by a textual span, table row or image-associated fact. Units are matched to the content actually delivered to the answerer. Visible historical summaries may contribute only the facts they contain; source identity alone does not import the original document's remaining content. The five-document limit controls selected documents, not the number of semantic units or all historical tokens. Local reference, loss and gain calculations use the same fixed state and rendering rule for every policy.

Panel selection and scoring. Both 500-question panels use fixed stratified selections. The MultiModalQA selection covers reasoning difficulty and text/image/table requirements; the WebQA validation selection covers factual and list questions. The laboratory maintains the MultiModalQA reference annotations and panel mapping, so the measurements are reported as local panel results rather than a reproduction of the public test evaluation. Development panels contain 64 and 56 disjoint questions. The score computation is word-based, with `nltk.word_tokenize`, lowercasing, stopword removal, repeated-token overlap, best-reference matching and zero for abstention. No alternative official leaderboard scorer is substituted.

B Feature Interfaces, Risk Training, and Execution Rules

B.1 Query and document representations

The frozen local classes are `BGE_M3_v2.0_Dense` and `CLIP_ViT_L14_336px`. Both configurations have hidden width 768. These are laboratory model specifications, not the default architectures of public checkpoints with similar names.

Text extraction. The text encoder returns the final layer's `last_hidden_state`, $H_T \in \mathbb{R}^{B\times L\times 768}$, where B is batch size and L is sequence length. For each item, the interface takes the token at index zero, `H_T[:,0,:]`, and L2-normalizes it. With $\mathcal{N}_2(w) = w/\|w\|_2$ for nonzero w, this gives $u(x_i) = \mathcal{N}_2(H_T[i, 0, :]) \in \mathbb{R}^{768}$. It does not average text tokens, truncate coordinates, or apply a projection. The text configuration has no built-in projection layer.

Image extraction. The 336×336 image encoder returns final-layer `last_hidden_state`, $H_I \in \mathbb{R}^{B\times(P+1)\times 768}$, with the class token at index zero and P patch tokens. It selects `H_I[:,1:,:]`, averages along the patch axis, and then L2-normalizes each row:

$$v_I = \mathcal{N}_2\left(\frac{1}{P}\sum_{p=1}^{P} H_I[:, p, :]\right) \in \mathbb{R}^{B\times 768}.$$

The class token is excluded. The local configuration records `projection_dim=768`; this extraction path uses the hidden states directly and applies no projection or other width-changing transformation.

Text input is the title and first 512 body characters; tables enter the text path. Up to three image features are averaged for a multi-image document before its visual block is L2-normalized. Missing modalities are represented by 0_{768}, not normalized. Parse or loading failures are excluded from the candidates, rather than treated as absent modalities. With these document blocks,

$$z_x = [u(x); v(x)] \in \mathbb{R}^{1536}, \qquad z_{q_t} = [u(q_t); 0_{768}] \quad \text{for a text-only query.} \tag{7}$$

Concatenation adds neither a projection nor whole-vector normalization. Its norm is $\sqrt{2}$ with two nonzero unit blocks and one with a single present block. Base cosine uses the full 1,536-dimensional vectors and their actual norms; it does not mask away missing modalities. The query uses the expanded current retrieval intent q_t. The risk input is $x_{td} = [u(q_t); z_d] \in \mathbb{R}^{2304}$, combining only the query text block with the complete document vector. Zero and non-finite vector handling follows Appendix B.3.

Table 4: Dimensions at the feature and risk interfaces. Missing modality blocks have the same width as present blocks.

Object	Dimension	Construction	Use
Text block	768	$u(x)$	CLS feature; present blocks L2-normalized
Visual block	768	$v(x)$	Patch mean, class token excluded; or zero block
Full query	1536	$[u(q_t); v(q_t)]$	Base cosine; visual block zero for text queries
Full document	1536	$[u(d); v(d)]$	Base cosine, alignment, novelty, risk input
Risk-head input	2304	$[u(q_t); z_d]$	Text-query block plus complete document vector

B.2 Risk supervision and training

A positive label marks document content adjudicated as erroneous or misleading with respect to the question and reference information. A negative label denotes content not established as erroneous under that review, including factually uncertain cases that were not confirmed incorrect after two-person review. Negative labels are therefore not certificates of source truth. Label construction uses rule-based screening, model-assisted preliminary annotations, and final human review against the document and reference answers. Generated-answer outcomes do not define these labels. The head consequently predicts this annotation convention, not the probability that a future generator will answer incorrectly.

The complete annotation pool contains 12,000 question–document pairs, disjoint from the evaluation panels. A class-stratified 80:20 split is performed at the pair level. Multiple documents associated with one question are assigned independently, so the validation split is not asserted to be question-disjoint from training. Checkpoint selection uses no independent test observations. Both splits have 36% positive labels; this is a property of the annotation pool, not an estimate of deployment prevalence.

Table 5: Risk-head annotation pool. Counts describe labelled pairs, not distinct questions.

Split	Positive	Negative	Total	Positive fraction
Training	3,456	6,144	9,600	36%
Validation	864	1,536	2,400	36%
Total	4,320	7,680	12,000	36%

For $x = x_{td}$, the network computes

$$\begin{aligned} h_1 &= D_{0.2}(\mathrm{ReLU}(W_1 x + a_1)), \\ h_2 &= D_{0.2}(\mathrm{ReLU}(W_2 h_1 + a_2)), \\ \ell &= W_3 h_2 + a_3, \qquad r = \sigma(\ell), \end{aligned} \tag{8}$$

where $W_1 \in \mathbb{R}^{256\times 2304}$, $W_2 \in \mathbb{R}^{128\times 256}$, and $W_3 \in \mathbb{R}^{1\times 128}$. All affine layers have biases. $D_{0.2}$ is dropout with probability 0.2 during training and the identity at inference. There are no residual connections or layer-normalization operations in this head. These dimensions imply 623,105 trainable head parameters, including biases; this count excludes the frozen encoders.

Training passes the raw logit ℓ to `BCEWithLogitsLoss`. Equivalently, the unweighted per-pair objective is

$$\mathcal{L}(\ell, y) = \mathrm{softplus}(\ell) - y\ell. \tag{9}$$

The sigmoid is not applied before this loss. It is applied once to the logit for inference-time scoring. The text and visual encoders are frozen throughout risk-head training; only the affine head parameters are updated. No class weights, sample weights, temperature scaling, or post-hoc probability calibration are used.

Table 6: Risk-head optimization and selection settings. Early stopping selects the checkpoint within a 30-epoch budget.

Item	Setting
Optimizer	Adam; learning rate 10^{-4}; weight decay 10^{-5}
Schedule	Cosine annealing
Batch size	128 labelled pairs
Maximum training	30 epochs
Early stopping	Five consecutive epochs without a decrease in validation loss
Checkpoint selection	Lowest validation loss
Random seed	42
Trainable modules	Risk head only; text and image encoders frozen
Loss	Unweighted binary cross-entropy with logits
Inference	Dropout off; one sigmoid; no additional calibration

B.3 Initialization, validity, and state updates

Before any evidence is admitted, $E_0 = \varnothing$. The initial rewritten query executes a plain retrieval with up to 50 candidates. The initial evidence selection ranks valid candidates by base cosine and takes at most five distinct documents. It does not evaluate gap, novelty, or risk. This first retrieval and evidence construction count toward the eight-cycle budget. The first ordinary controlled admission uses the initialized plain evidence as its reference.

Subsequent E_t is the accumulated, document-ID-deduplicated memory of admitted sources used to compute the mean and novelty maximum. It is distinct from the at-most-five new identifiers

returned by an admission call. Historical summaries contribute to evaluation only when their content is actually delivered to the answerer. An empty admission result preserves the existing evidence and triggers repair when the termination rules permit another retrieval; rejected positions are not automatically refilled.

Table 7: Validity and fallback rules. Numerical invalidity and a large learned risk are different cases.

Condition	Selection action
Absent modality	Supply a zero block; a usable nonzero complete vector participates in the ordinary calculations.
Parse or loading failure	Remove the failed candidate from the working list.
Unavailable risk estimate	Set $r = 1$ and retain the otherwise valid candidate for scoring.
Empty evidence pool	Initialize or fall back to base-cosine ranking without gap, novelty, or risk.
Zero evidence mean	Use base-score ranking; do not divide by its zero norm.
Constant candidate gap	Use base ranking instead of dividing by a zero gap range.
Non-finite admission score	Mark the candidate ineligible for final Top-5 selection; do not compare NaN or infinity as an ordinary score.
Empty selection	Preserve existing evidence; repair only within the remaining stopping budget.
Changed candidate pool	Recompute pool extrema, scores, duplicate resolution, and ranking on the current valid records.

A missing modality is not itself a zero complete vector: $[u; 0]$ remains nonzero when $u \neq 0$. A full-vector zero norm makes the required cosine undefined and cannot supply a finite ordinary score; such a record is not eligible for final selection. Risk failure alone instead receives the finite value one. At $r = 1$, the term $0.25(1 - r)$ is zero, but the other terms can still make the candidate eligible. This is a maximal penalty in that term, not a hard veto.

For the finite-score path, the raw base cosine lies in $[-1, 1]$, gap and novelty in $[0, 2]$, and normalized gap in $[0, 1]$. The mean vector is computed before taking its cosine. Scoring occurs before document-level deduplication; for duplicate identifiers, keep the greatest admission score and then greatest base score. Eligible identifiers are sorted by decreasing admission score, decreasing base score, and increasing ID. An update that removes unusable records changes the pool over which the gap extrema are recomputed.

The score is not a probability. Write $w = 0.20b + 0.30(1 - \widetilde{g}) + 0.25n$. The threshold condition is

$$s \geq 0.65 \quad \Longleftrightarrow \quad r \leq 4w - 1.6. \tag{10}$$

If $w < 0.40$, no $r \in [0, 1]$ reaches the threshold; if $w \geq 0.65$, every such risk value does. Increasing only risk changes the score by $-0.25\Delta r$. A candidate can remain eligible yet lose a top-five position, and pool-dependent gap normalization can change the alignment of otherwise unchanged records. These are numerical properties, not empirical calibration or truth guarantees.

B.4 Sufficiency, online novelty, and termination priority

The sufficiency module is an instruction-guided LLM evaluator, separate from the binary risk head. It receives the question, admitted-evidence summaries, and a candidate draft, and returns a self-assessed score $a \in [0, 100]$. The parsed score is $\widehat{c} = a/100$, without sigmoid or reference-support matching. Out-of-range values, malformed responses, and timeouts are non-passing assessments with no imputed score. The threshold 0.65 applies to this self-report, not an observed support fraction or calibrated correctness probability.

The online span inventory is separate from the offline reference support S_i. After document-ID deduplication, newly admitted text is split at full stops and semicolons and exact-matched against previously accepted spans. Differently worded expressions of the same fact can count as new strings; this update performs no semantic-entailment or factual-correctness test.

Let $\mathcal{F}_t$ denote this inventory and B_t the spans supplied by the current admissions. The monitoring rule is

$$\Delta_t = |B_t \setminus \mathcal{F}_{t-1}|, \quad \mathcal{F}_t = \mathcal{F}_{t-1} \cup B_t, \quad k_t = \begin{cases} k_{t-1} + 1, & \Delta_t = 0, \\ 0, & \Delta_t > 0, \end{cases} \qquad k_0 = 0. \tag{11}$$

The counter updates at each retrieval-cycle end. Two consecutive cycles without a new string produce $0 \to 1 \to 2$ and trigger stagnation.

Citation validation requires a source reference for every factual claim and rejects malformed citations, out-of-range identifiers, and references without a delivered source. It checks source linkage, not semantic entailment. A failing draft permits at most two rewrites without additional retrieval, with sufficiency re-evaluated after each rewrite.

Fallback eligibility. A stored draft is eligible when its answer text is nonempty, parseable, and not an abstention marker. Passing citations and a usable sufficiency assessment are not required: missing, malformed, timed-out, or out-of-range assessments do not exclude an otherwise eligible draft. An empty eligible set yields refusal. Eligibility establishes output availability, not factual correctness.

Deterministic fallback selection. Let $\mathcal{A}$ be the eligible drafts, and let $\mathcal{A}_+$ contain those with a finite, parsed sufficiency score $\widehat{c}(a) \in [0, 1]$. A draft without such a score remains eligible but cannot displace a scored draft. If $\mathcal{A}_+ \neq \varnothing$, define

$$c_* = \max_{a \in \mathcal{A}_+} \widehat{c}(a), \qquad \mathcal{B} = \{a \in \mathcal{A}_+ : \widehat{c}(a) \geq c_* - 0.05\}.$$

Only $\mathcal{B}$ enters the next comparison. Scores inside this band are not compared again. This maximum-relative band, rather than pairwise approximate comparisons, defines the near-tie rule.

For draft a, let $j(a)$ count citation matches, $n_{\mathrm{ref}}(a)$ count referenced documents, $\ell(a)$ be its answer-text character count, and $\mathrm{id}(a)$ be its unique draft identifier. Its retrieval relevance is

$$\rho(a) = \begin{cases} \max_{d \in \mathcal{C}(a)} b_a(d), & \mathcal{C}(a) \neq \varnothing, \\ 0, & \mathcal{C}(a) = \varnothing, \end{cases}$$

where $\mathcal{C}(a)$ is the referenced-document collection and $b_a(d)$ is the query–document cosine associated with the draft. Selection within $\mathcal{B}$ maximizes citation matches, then retrieval relevance, then character count, with increasing draft ID as the last key. When every eligible draft is unscored, selection instead maximizes retrieval relevance, then the number of referenced documents, and finally uses increasing draft ID. Relevance is compared numerically: an uncited draft with value zero can precede a cited draft with negative cosine. Multiple drafts referring to the same document remain distinct.

Algorithm 2 Selection of one fallback draft

Require: Stored drafts with answer text, parsed sufficiency, citations, relevance, and unique draft IDs.
Ensure: One eligible draft, or the refusal marker.

$\mathcal{A} \leftarrow$ nonempty, parseable, non-abstaining drafts.
if $\mathcal{A} = \varnothing$ **then**
    return refusal.
end if
$\mathcal{A}_+ \leftarrow \{a \in \mathcal{A} : \widehat{c}(a)$ is finite and in $[0, 1]\}$.
if $\mathcal{A}_+ \neq \varnothing$ **then**
    $c_* \leftarrow \max_{a \in \mathcal{A}_+} \widehat{c}(a)$.
    $\mathcal{B} \leftarrow \{a \in \mathcal{A}_+ : \widehat{c}(a) \geq c_* - 0.05\}$.
    return the lexicographic minimum under $(-j(a), -\rho(a), -\ell(a), \mathrm{id}(a))$ over $\mathcal{B}$.
else
    return the lexicographic minimum under $(-\rho(a), -n_{\mathrm{ref}}(a), \mathrm{id}(a))$ over $\mathcal{A}$.
end if

Algorithm 2 specifies one returned draft, not a pairwise total ranking of all unused drafts. Missing sufficiency is not imputed as zero. Fallback eligibility and selection remain distinct from passing

the normal sufficiency and citation checks. Table 8 determines when this selector is invoked; the fixed-state evidence-selection comparison does not invoke it.

Table 8: Stagnation-refusal policy, evaluated from the highest-priority applicable condition downward. Draft fallback is not equivalent to passing verification.

Condition	Action
Eight retrieval cycles reached	Stop retrieval and apply Algorithm 2, regardless of the early-answer checks; refuse if none is eligible.
Two-cycle stagnation before budget exhaustion	Stop retrieval and return the refusal marker.
Sufficiency ≥ 0.65 and citations pass	Output the current answer and terminate normally.
Citation check fails with attempts remaining	Rewrite the draft, re-evaluate sufficiency and citations, with at most two rewrites.
Two citation rewrites exhausted without passage	Stop rewriting and apply Algorithm 2; refuse if none is eligible.
Insufficient evidence, no higher-priority stop	Repair the retrieval intent and continue within the retrieval-cycle limit.
Invalid assessment	Do not impute a sufficiency score. The applicable budget, stagnation, and draft-eligibility rules still govern termination.

At simultaneous budget exhaustion and stagnation, the budget rule applies and an eligible draft may be returned. The reported full-system scores include such fallback outputs; they are not restricted to answers passing the normal verification gate. The answer-set partition does not identify which termination paths supply the gains. By contrast, the local comparison holds history, candidates, and scoring records fixed and changes only admission.

C Held-Out Evidence Ranking and Paired Outcomes

The held-out panel contains 300 MultiModalQA and 300 WebQA questions, with 283 and 279 usable first-control states. Question identities are disjoint from training, development, and the system-evaluation panels. The controller logic is held fixed across selectors. The analysis is conditional on the 562 reached states.

Each selector receives the same recorded candidate pool, representations, history, and renderer, with up to five selected documents. Full–Rank retains Eq. (4) but omits the 0.65 admission threshold. The no-risk score is

$$s_{-r}(d) = 0.20b(d) + 0.30(1 - \widetilde{g}(d)) + 0.25n(d). \tag{12}$$

No remaining coefficient or feature normalization is changed. A common positive rescaling by $1/0.75$ would not affect this threshold-free order, but is not applied. These variants do not replace the definition of the thresholded system.

MMR greedily maximizes

$$m_\lambda(d \mid A) = \lambda b(d) - (1 - \lambda) \max_{j \in A} \cos(z_d, z_j), \tag{13}$$

where A grows within the current selection call. It is not the cross-turn memory E_t used by Full–Rank. The first document has highest base similarity; subsequent ties use decreasing base similarity and increasing document ID. The development setting is $\lambda = 0.6$, selected from $\{0, 0.1, \ldots, 1\}$ using 50 disjoint states, 25 from each dataset. The selection priorities are higher CE, fewer loss states, more retained support, and then a smaller parameter. This value is fixed for held-out evaluation.

Table 9: Held-out first-control outcomes. CE and loss count questions; V counts visible reference-support units.

Panel	Selector	N	CE	Loss	V
MMQA	MMR	283	118	92	172
MMQA	No risk	283	129	86	188
MMQA	Full–Rank	283	134	83	195
WebQA	MMR	279	121	97	216
WebQA	No risk	279	132	91	235
WebQA	Full–Rank	279	138	88	242

C.1 Paired cells and the four-test family

The first member of every pair is Full–Rank. Cells are ordered $(n_{11}, n_{10}, n_{01}, n_{00})$. Label 1 means complete evidence for CE, and at least one missing initial support identity for loss. The two marginal event counts are $n_{11}+n_{10}$ and $n_{11}+n_{01}$. For a binary difference $d_i = y_i^F - y_i^B$, its sum is $n_{10}-n_{01}$ and its squared sum is $n_{10}+n_{01}$.

Table 10: All dataset-level paired cells for the held-out ranking comparisons.

Panel	Comparator	Outcome	n_{11}	n_{10}	n_{01}	n_{00}
MMQA	MMR	CE	106	28	12	137
WebQA	MMR	CE	109	29	12	129
MMQA	MMR	Loss	70	13	22	178
WebQA	MMR	Loss	74	14	23	168
MMQA	No risk	CE	127	7	2	147
WebQA	No risk	CE	130	8	2	139
MMQA	No risk	Loss	75	8	11	189
WebQA	No risk	Loss	79	9	12	179

For $b = n_{10}$ and $c = n_{01}$, the two-sided exact McNemar probability under independent question pairs and equal marginal event probabilities is

$$p = \min\left\{1, \frac{2}{2^{b+c}} \sum_{k=0}^{\min(b,c)} \binom{b+c}{k}\right\}, \quad p = 1 \text{ if } b+c = 0. \tag{14}$$

The four null hypotheses compare Full–Rank with MMR and no risk for CE and loss, respectively. All four remain in the retrospective family. If $p_{(1)} \le \cdots \le p_{(4)}$, the adjusted probability is

$$p_{(j)}^{\text{Holm}} = \min\left\{1, \max_{1\le k\le j}(5-k)p_{(k)}\right\}. \tag{15}$$

Adjustment uses unrounded probabilities. The endpoint diagnostic is analyzed separately from this family.

Table 11: Pooled cells and complete retrospective family. p and Holm p assume independent question pairs and do not incorporate source-cluster dependence.

Comparator	Outcome	n_{11}	n_{10}	n_{01}	n_{00}	p	Holm p
MMR	CE	215	57	24	266	0.000317	0.001269
MMR	Loss	144	27	45	346	0.044371	0.088742
No risk	CE	257	15	4	286	0.019211	0.057632
No risk	Loss	154	17	23	368	0.429591	0.429591

The CE contrast with MMR survives this four-test correction. The no-risk CE contrast does not: $p = 0.0192108$ becomes 0.0576324. These calculations describe the independent-question model;

they are not source-adjusted tests. Retained-support differences of 49 and 14 units are descriptive totals.

D SEVEN-SYSTEM EVALUATION

The two panels contain 500 questions each. Answer-score sums use $f_i \in [0, 1]$ and are recorded to two decimal places; dividing by 500 and multiplying by 100 yields a dataset F1 percentage. Displayed decimal precision is not an estimate of replication uncertainty. Reference support is question-indexed: a source reused by two questions does not merge their separate evidence requirements.

Table 12: Dataset-level results for the principal comparisons. Each panel contains 500 questions; all values are percentages.

Panel	Configuration	CE (%)	MR (%)	F1-all (%)
MultiModalQA	Single BM25	19.60	18.00	15.246
MultiModalQA	Agentic Hybrid	37.80	34.33	18.608
MultiModalQA	EGAL-SC	43.60	39.87	20.194
WebQA	Single BM25	23.80	19.29	17.034
WebQA	Agentic Hybrid	41.60	34.80	20.432
WebQA	EGAL-SC	45.20	39.91	23.672

Table 13: MultiModalQA panel: disclosed aggregate inputs. $N = 500$; $T = 1372$ support units. F is the sum of per-question F1 values on a zero-to-one scale; C, A, and R denote complete-evidence questions, answered questions, and recovered support units.

Configuration	F	C	A	R
Single BM25	76.23	98	212	247
Single Hybrid	82.18	126	195	312
Agent-BM25	89.72	172	152	429
Agentic Hybrid	93.04	189	153	471
Text-only LogicalRAG	56.87	107	119	268
Multimodal LogicalRAG	49.32	122	96	309
EGAL-SC	100.97	218	207	547

Table 14: WebQA panel: disclosed aggregate inputs. $N = 500$; $T = 1586$ support units. F is the sum of per-question F1 values on a zero-to-one scale; C, A, and R denote complete-evidence questions, answered questions, and recovered support units.

Configuration	F	C	A	R
Single BM25	85.17	119	244	306
Single Hybrid	92.74	143	226	384
Agent-BM25	98.43	190	177	501
Agentic Hybrid	102.16	208	177	552
Text-only LogicalRAG	62.73	121	139	317
Multimodal LogicalRAG	54.04	132	112	359
EGAL-SC	118.36	226	238	633

For EGAL-SC, the score sum is $100.97 + 118.36 = 219.33$ and the answered count is $207 + 238 = 445$, giving $\mathrm{F1}_{\mathrm{all}} = 21.933\%$ and $\mathrm{F1}_{\mathrm{answered}} = 49.28764\%$. The corresponding Hybrid values are 195.20, 330, 19.520%, and 59.15152%.

The pooled support recall is

$$\mathrm{MR}_{\mathrm{pooled}} = 100\frac{R_{\mathrm{MMQA}} + R_{\mathrm{WebQA}}}{1372 + 1586}. \tag{16}$$

It is 39.89182% for EGAL-SC and 34.58418% for Agentic Hybrid, whereas the dataset-macro values are 39.89027% and 34.56699%. Complete evidence weights questions equally; micro recall weights individual support units. Their aggregate percentages therefore need not have a fixed ordering.

Table 15: Overall results on the two 500-question panels. All values are percentages. CE denotes complete presented evidence; MR is the dataset-macro support-unit recall. F1-all includes abstentions as zero; F1-ans. is evaluated on each method's answered subset. Coverage is the fraction of questions answered.

Configuration	CE	MR	F1-all	Coverage	F1-ans.
Single BM25	21.70	18.65	16.140	45.60	35.39
Single Hybrid	26.90	23.48	17.492	42.10	41.55
Agent-BM25	36.20	31.43	18.815	32.90	57.19
Agentic Hybrid	39.70	34.57	19.520	33.00	59.15
Text-only LogicalRAG	22.80	19.76	11.960	25.80	46.36
Multimodal LogicalRAG	25.40	22.58	10.336	20.80	49.69
EGAL-SC	44.40	39.89	21.933	44.50	49.29

Table 16: Warm-start end-to-end latency and token use on 1,000 questions per configuration. Mean and P95 are in seconds; tokens include prompt and completion tokens. Shared upper budgets do not imply equal realized computation.

Configuration	Mean (s)	P95 (s)	Tokens/question
Single BM25	47.23	168.45	1,410
Single Hybrid	56.87	215.32	2,087
Agentic Hybrid	365.42	1,062.78	15,835
Multimodal LogicalRAG	412.56	1,287.34	19,718
EGAL-SC	485.62	1,025.37	22,011

E Observed Answer-Set Decomposition

The four answer sets partition the 500 questions in each panel. In this section J denotes joint answers, E EGAL-only answers, H Hybrid-only answers, and Z joint abstentions. A score sum cannot exceed its group's question count because every per-question F1 lies in $[0, 1]$.

Table 17: Answer-set decomposition on 1,000 questions. Score sums use per-question F1 in $[0, 1]$; the final column gives each set's contribution to the all-question F1 difference in percentage points.

Answer set	N	EGAL sum	Hybrid sum	Contribution
Both answer	307	171.50	180.20	-0.870
EGAL only	138	47.83	0.00	$+4.783$
Hybrid only	23	0.00	15.00	-1.500
Neither answers	532	0.00	0.00	0.000
Total	1,000	219.33	195.20	$+2.413$

Table 18: Per-panel answer-set counts and score sums. Scores are summed before converting to percentages.

Panel	Set	N	EGAL sum	Hybrid sum
MMQA	J	142	79.32	85.47
MMQA	E	65	21.65	0.00
MMQA	H	11	0.00	7.57
MMQA	Z	282	0.00	0.00
WebQA	J	165	92.18	94.73
WebQA	E	73	26.18	0.00
WebQA	H	12	0.00	7.43
WebQA	Z	250	0.00	0.00

The groups recover the answered counts: EGAL has $142 + 65 = 207$ MMQA and $165 + 73 = 238$ WebQA answers; Hybrid has $142+11 = 153$ and $165+12 = 177$. Score sums remain 100.97/118.36 for EGAL and 93.04/102.16 for Hybrid.

Table 17 gives the pooled decomposition. Across the 307 common answers, EGAL and Hybrid sums are 171.50 and 180.20. Their conditional F1 values are 55.86319% and 58.69707%, a difference of -2.83388 percentage points. The nonoverlapping sets add $47.83 - 15.00 = 32.83$ score units, whereas the common set contributes -8.70, yielding 24.13 score units or 2.413 all-question percentage points.

Table 19: Paired moments on jointly answered questions only, with $d_i = f_i^E - f_i^H$.

Panel	N_J	$\sum d_i$	$\sum d_i^2$
MultiModalQA	142	-6.15	8.72
WebQA	165	-2.55	5.14

For panel d, let $S_d = \sum_i d_i$, $Q_d = \sum_i d_i^2$, N_d be the jointly answered count, $N = \sum_d N_d$, and $w_d = N_d/N$. The sample variance is $s_d^2 = (Q_d - S_d^2/N_d)/(N_d - 1)$, giving

$$\widehat{\Delta} = \sum_d w_d \frac{S_d}{N_d}, \qquad \widehat{\mathrm{Var}}(\widehat{\Delta}) = \sum_d w_d^2 \frac{s_d^2}{N_d}. \tag{17}$$

Writing $v_d = w_d^2 s_d^2/N_d$, the interval is $100[\widehat{\Delta} \pm t_{\nu,0.975}\sqrt{\sum_d v_d}]$, with $\nu = \frac{(\sum_d v_d)^2}{\sum_d v_d^2/(N_d-1)}$.

Using Eq. (17) with weights $142/307$ and $165/307$ gives standard error 1.20324 percentage points and an approximate 95% interval $[-5.20257, -0.46518]$ for the common-answer difference. This is conditional on the observed shared answer set and independent questions; it is not an interval for the full 1,000-question F1 difference and does not identify a causal response to abstaining or answering.

F Full-Trajectory Support-Loss Comparison

The trajectory comparison concerns 912 questions with typed controls. The initial plain reference is shared across Direct and guarded configurations, and each question contributes one final loss label. Several calls in one trajectory do not create independent questions. The state-level loss definition in Eq. (5) instead compares one selected context with its Plain-5 presentation.

Table 20: Full-trajectory support-loss transitions. These cells are distinct from the fixed-state counts.

	Guarded loss	Guarded no loss	Total
Direct loss	25	63	88
Direct no loss	5	819	824
Total	30	882	912

There are $63 - 5 = 88 - 30 = 58$ fewer endpoint losses; the relative reduction is $1 - 30/88 = 65.90909\%$. Of the 88 Direct-loss questions, 63 have no guarded loss (71.59%); of 824 Direct-no-loss questions, five have guarded loss (0.61%). These different conditional denominators are not added.

Conditioning on 68 discordant pairs under equal marginal loss probabilities and independent question pairs gives

$$p = \frac{2}{2^{68}} \sum_{k=0}^{5} \binom{68}{k} = 7.6510629429 \times 10^{-14}. \tag{18}$$

This retrospective McNemar diagnostic is separate from the held-out four-test family and is used for the complete-trajectory comparison. It does not adjust source clustering and does not isolate admission from repair and verification. Also, the 556 questions without complete final evidence $(1000 - 444)$ are not the same event as the 30 that lose initial support.

G Human Assessment and Label Agreement

The human sample contains 400 outputs (200 per dataset), stratified by method and answer/refusal and enriched for difficult cases. Assessors viewed shuffled outputs without system names; a third assessor adjudicated disagreements. Binary correctness requires all key facts to be supported and free of errors. Partially correct outputs or unsupported answers are labelled 0; an appropriate refusal is labelled 1. This is a correctness convention that includes justified abstention, not a measure of answering accuracy alone.

Table 21: Output-level binary agreement. Counts are ordered $(1,1), (1,0), (0,1), (0,0)$.

Comparison	n_{11}	n_{10}	n_{01}	n_{00}	Agreement	κ
Two assessors	172	28	26	174	86.50%	0.730
Judge vs adjudication	158	19	22	201	89.75%	0.793

With $p_o = (n_{11} + n_{00})/N$ and expected agreement p_e obtained from the marginal label frequencies, $\kappa = (p_o - p_e)/(1 - p_e)$. The assessor table uses pre-adjudication labels; the judge table uses the adjudicated reference. Agreement measures consistency under this rubric. It does not establish population error rates, independent source truth or a ranking of methods.

H Fixed-Answer Coverage Analysis

This analysis retains subsets of existing answers without regenerating text. Its inputs are the answer count a_d and score sum F_d for each dataset.

Bounds for retaining a fixed number of answers. If scores lie in $[0, 1]$ and $k_d \leq a_d$ existing answers are retained, their retained score sum W_d satisfies

$$\max\{0, F_d - (a_d - k_d)\} \leq W_d \leq \min\{F_d, k_d\}. \tag{19}$$

Removing $a_d - k_d$ scores removes at most that much mass; retained mass cannot exceed F_d or k_d. Allocating mass preferentially to removed or retained answers attains the respective endpoints under the marginal constraints. These bounds remain valid with the observed answer partition but need not

be sharp conditional on its additional moments. They are not confidence intervals or guarantees for a deployable selector.

To match Agentic Hybrid's dataset-specific answer counts, take $(k_{\mathrm{MMQA}}, k_{\mathrm{WebQA}}) = (153, 177)$. For EGAL-SC, (a_d, F_d) are $(207, 100.97)$ and $(238, 118.36)$. Equation (19) gives retained sums in $[46.97, 100.97]$ and $[57.36, 118.36]$. Their total is $[104.33, 219.33]$, yielding all-question F1 in $[10.433, 21.933]\%$ on the original 1,000 questions. Compared with Hybrid's 19.520%, the aggregate-compatible difference lies in $[-9.087, 2.413]$ percentage points. Its sign is not identified by the marginal scores.

Uniform random thinning. Uniformly select exactly k_d of the a_d existing answers in each dataset, independently of their scores. Each answer is retained with probability k_d/a_d, hence linearity of expectation gives

$$\mathbb{E}[\mathrm{F1}_{\mathrm{all,thin}}] = \frac{100}{1000}\left(\frac{153}{207}\,100.97 + \frac{177}{238}\,118.36\right) = 16.265403\%. \tag{20}$$

This exact random-thinning expectation matches per-dataset answer counts, not Hybrid's answered set. It is not a new model run and does not reduce inference already performed.

Exceeding Hybrid after removing 115 EGAL answers requires removed score mass below 24.13, or a removed-answer mean below $24.13/115 = 0.209826$. The summaries do not establish an online rule that identifies such answers; the random-thinning expectation likewise does not imply that every alternative selection is inferior.