SABR: Structure-Aware Budgeted Retention under Fixed Token Budgets

Anonymous authors
Paper under double-blind review

ABSTRACT

Selecting relevant fragments is not sufficient when a context budget separates evidence needed together. We introduce Structure-Aware Budgeted Retention (SABR), a training-free selector combining frozen semantic scores, local coordination, and exact generator-token packing. We compare SABR with a strong frozen reranker at identical per-question context and prompt lengths, separating selection quality from full-context quality and online cost. On a previously inspected reserve of 1,384 questions across four tasks and three generators, SABR improves SQuAD F1 at a 70% context cap by 3.18, 4.19, and 4.23 points; the latter two paired 95% intervals exclude zero. The four-task macro difference is $+0.92$ $[-0.08,\ 1.91]$. A separate 896-question cohort yields a positive SQuAD estimate but does not confirm an aggregate gain. A matched-token sentence-unit pilot increases answer-string presence on SQuAD and mean intact-support fraction on HotpotQA by 10.94 and 13.80 percentage points. Timing studies show that exact implementation optimizations reduce selection overhead, but do not establish an overall speedup over full context. These findings identify a concrete regime in which coordinated retention improves fixed-budget selection, while distinguishing retained evidence, answer quality, and computational efficiency.

1 INTRODUCTION

A number without its unit, an entity without its qualifier, or a statement without its temporal scope can be relevant to a question yet insufficient to answer it. Under a fixed context budget, fragments compete for space even when their value depends on which other fragments survive. Context selection therefore requires more than ranking individual relevance: it also requires deciding *what should be retained together*.

This distinction motivates a controlled question: can coordinated selection improve answers beyond strong relevance ranking *without keeping more input*? Prompt compression and context-aware extraction already address related efficiency problems (Jiang et al., 2023; 2024; Pan et al., 2024; Xu et al., 2024; Hwang et al., 2025). Our focus is narrower. The supplied context and generator remain fixed, and both compressed arms have identical realized context and prompt token counts. This removes retained length as an explanation for an observed quality difference.

SABR combines frozen bi-encoder and cross-encoder relevance with deterministic local propagation, bounded entity/proximity bonuses, and group-aware packing. It preserves source order and enforces the budget with the generator's tokenizer. Selection uses no reference answers, supporting-fact labels, or dataset identifiers. The resulting method is inspectable and requires no task-specific training, although its rules were developed using previously inspected data.

Our contribution is a concrete selector and an evaluation that separates three questions: whether selected evidence is more useful at the same input length, how much quality is lost relative to full context, and whether shorter inputs offset selection overhead. Cross-generator comparisons, a separate confirmation cohort, representation and structure diagnostics, and timing studies address these questions without pooling incompatible protocols. The evidence supports a task-dependent selection benefit, rather than a universal ordering of compressors.

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.

2 Related Work and Decision Problem

2.1 Selection after retrieval

Retrieval chooses information from a corpus; context selection decides what a generator receives from an already supplied input. Dense passage retrieval and retrieval-augmented generation address the former problem (Karpukhin et al., 2020; Lewis et al., 2020). SABR addresses the latter. It reuses established bi-encoder and cross-encoder models rather than introducing a new relevance model (Reimers and Gurevych, 2019; Beijing Academy of Artificial Intelligence, 2023b;a).

LLMLingua and LongLLMLingua use language-model signals for prompt compression, while LLMLingua-2 learns a distilled token classifier (Jiang et al., 2023; 2024; Pan et al., 2024). RECOMP and EXIT study compression and context-aware extraction for retrieval-augmented generation (Xu et al., 2024; Hwang et al., 2025). SABR instead coordinates retention using frozen semantic scores and observable input structure. Its primary comparison is with cross-encoder ranking under exactly matched realized lengths. A separate comparison with a token-cap-adapted learned compressor is reported in Appendix F, not substituted for this control.

Context structure is broader than adjacency. Evidence position can change model behavior (Liu et al., 2024), and LongBench evaluates diverse long-context tasks (Bai et al., 2024). Restoring source order does not preserve every original position after deletion. Adaptive-RAG and Self-RAG additionally change retrieval or generation behavior (Jeong et al., 2024; Asai et al., 2023); SABR's deterministic input route does neither. Its optional escalation wrapper is analyzed separately in Appendix G.

2.2 Budget, matching, and evaluation targets

For question q, ordered fragments $C = (c_1, \ldots, c_m)$, generator G, and its tokenizer count τ_G, let C_S denote selected fragments rendered in source order. At context cap ρ, require

$$\tau_G(C_S) \le B(C, \rho), \qquad B(C, \rho) = \max\{1, \lfloor \rho \tau_G(C) \rfloor\}. \tag{1}$$

The denominator includes supplied context only. Questions, instructions, and chat-template tokens are outside it, but the complete prompt must satisfy the generator's input limit. A 70% cap means retaining *at most* 70% of context tokens, not a 70% reduction or a 30% saving on the entire request. Unused budget is allowed.

Let $P(q, C)$ be the rendered prompt and $F(G(P), y) \in [0, 100]$ the task-specific answer score. The primary SABR–reranker comparison enforces

$$\tau_G(C_S) = \tau_G(C_R), \qquad \tau_G(P(q, C_S)) = \tau_G(P(q, C_R)). \tag{2}$$

Define $\Delta_{\text{sel}} = \mathbb{E}[F_S - F_R]$ and $\Delta_{\text{full}} = \mathbb{E}[F_S - F_{\text{full}}]$. The first measures selection value at matched length; the second measures the quality cost of compression. Positive Δ_{sel} can coexist with negative Δ_{full}. Runtime is a third endpoint because constructing a shorter input itself requires computation.

Complementarity and opportunity cost. Preserving a neighboring qualifier may make a retained fact usable, but at a fixed budget it displaces something else. Adjacency and entity overlap are therefore proxies for useful relationships, not guarantees of complementary evidence. SABR's comparison with reranking tests the combined semantic and coordination policy. It does not by itself identify a causal contribution of structure; the narrower controls in Section 6 address that distinction.

3 SABR: Coordinating Evidence before Packing

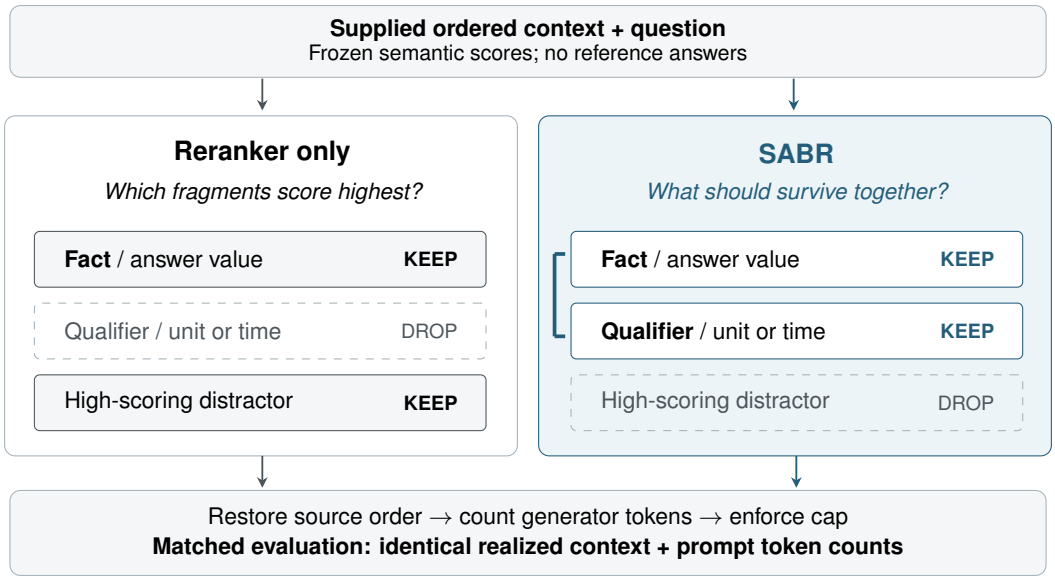

Figure 1: Different evidence at the same realized input length. SABR may retain a related qualifier by displacing another fragment. The illustration is constructed, not a benchmark outcome; box sizes are not token costs. Both compressed arms restore source order and pass the context/prompt-token audit.

3.1 Shared representation and frozen relevance

All primary arms use the same supplied representation. Prose is split at sentence and clause boundaries; long fragments are divided into units of at most eight whitespace-separated words. Document/title information and ordering are retained. This fine granularity makes local continuity explicit, but can split relations that sentence-level selection would already preserve.

Let u_i and v_i be within-context normalized bi-encoder and cross-encoder scores. A deterministic route $r(q, C)$ distinguishes multi-document, boolean, arithmetic, and remaining factoid inputs using only observable text. The semantic score is

$$s_i = \alpha_r u_i + (1 - \alpha_r) v_i, \tag{3}$$

with $\alpha_r = 0.20, 0.85, 0.75, 0.35$ for those routes, respectively. Query-term, entity-like, title, and boundary signals supplement this score. Helpers and exact coefficients appear in Appendix A.

3.2 Coordination and exact token enforcement

Boolean and multi-document inputs use adaptive greedy packing with bounded overlap and proximity bonuses and a redundancy penalty. Other inputs with fewer than 25% digit-containing fragments first propagate semantic relevance within three positions, $\widetilde{s}_i = \max_{|j-i|\leq 3} s_j$, then use adaptive packing. Remaining high-numeric-density inputs prioritize consecutive sentence groups, followed by fragment-level fill. These branches encode hypotheses about continuity and numerical relations, not learned guarantees.

An internal cost estimate proposes a subset. SABR restores source order, renders it, and checks actual generator tokens. An over-budget proposal is retried with a reduced internal budget, for at most twenty attempts; direct-token pruning enforces the cap afterward. Ties favor earlier fragments. Reranker-only uses the same frozen cross-encoder but not SABR's semantic mixture or coordination. Matching additionally enforces Equation 2; it controls retained length, not helper cost. The primary mean retention is 49.82–49.99% at the 50% cap and 69.66–69.98% at 70%.

4 EXPERIMENTAL DESIGN

4.1 COHORTS AND THEIR DISTINCT ROLES

The primary reserve contains the first 346 commonly eligible questions per task after intersecting three generators' 4,096-token input limits: 1,384 questions from BoolQ, SQuAD, HotpotQA, and a restricted DROP subset. These public training examples were inspected during method development. The primary estimate is therefore a development-reserve result, not an untouched test estimate.

The confirmation cohort fixes 256 previously unused public-validation questions each from BoolQ, SQuAD, and HotpotQA, and 128 from DROP, before generation. Its six arms are full context, reranker, SABR, and removals of semantic scoring, entity bridging, or the boundary prior. "Independent" denotes unused questions in the recorded SABR experiments, not certified absence from checkpoint pretraining. Identical final inputs reuse cached answers without being counted as additional independent observations.

Table 1: Evaluation inventory. Counts are question/context units, not generation calls. Different cohorts, scoring rules, and matching protocols are not pooled.

Study	Units	Generator and purpose
Primary reserve	1,384	Three generators; 50/70%; exact-length quality
Independent confirmation	896	Qwen 7B; 70%; new cohort and components
Sentence-unit pilot	128	Qwen 7B; matched tokens; retained evidence
Matched-prefix diagnostic	1,498	Mistral 7B; 70%; no-structure control
Native-length timing	12	Qwen 7B; 50/70%; fixed output length
Engineering retest	32	Mistral 7B; original/incremental implementations
Long-input challenge	38	Mistral 7B; actual full prompts 4,105–6,986 tokens

The sentence pilot uses 64 new SQuAD and 64 new HotpotQA examples. The matched-prefix diagnostic is a separate Mistral reserve, not another confirmation cohort. The 32-context engineering set is length-stratified and train-derived. The long-input challenge consists of 38 HotpotQA contexts excluded by the earlier 4,096-token limit; its input limit is raised to 8,192. The smaller timing, routing, and calibration studies do not enlarge the primary sample size.

4.2 MODELS, METRICS, AND PAIRED INFERENCE

Primary generators are Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, and Mistral-7B-Instruct-v0.3 (Qwen Team, 2024; Mistral AI, 2024). Quality evaluation uses BF16, SDPA attention, batch-one greedy generation, and at most 64 new tokens. All selectors see the same question and supplied context. The sentence pilot separately uses FP16; the twelve-context timing study fixes outputs to 32 tokens.

A common gold-blind answer extractor precedes task scoring. BoolQ uses yes/no accuracy (Clark et al., 2019); SQuAD uses v1.1 answer F1 (Rajpurkar et al., 2016); HotpotQA uses answer F1, not supporting-fact or joint scores (Yang et al., 2018). DROP covers only the study's single-span subset, not the complete official benchmark (Dua et al., 2019). The primary macro is an equal-weight average of task-specific 0–100 scores, not uniform accuracy. Audited rescoring changes 2,475 of 10,380 stored SQuAD/HotpotQA answers relative to the earlier generic-overlap scorer. All primary claims use the audited values.

Primary intervals use 10,000 paired question bootstraps. Macro resampling is stratified by task and preserves each question's three model outputs together; generators are not independent data replications. Sign-flip tests use 10,000 replicates and Holm adjustment within stated families. Task-level intervals are pointwise and do not account for method-development uncertainty. Separate diagnostics retain their own units and analysis protocols (Appendices B–E).

5 Answer Quality at Matched Token Counts

5.1 A repeated SQuAD benefit at the 70% cap

Table 2 shows the clearest primary result: SABR gains 3.18, 4.19, and 4.23 SQuAD F1 over reranking at identical realized context and prompt lengths. The latter two paired intervals exclude zero. The direction repeats across generators, but the questions are shared, so this is not three independent replications. Qwen models recover part of the quality lost through reranking; Mistral's point estimate slightly exceeds full context.

Table 2: Primary SQuAD at 70% ($n = 346$ shared questions). Differences are computed before rounding; intervals are paired, pointwise 95%.

Generator	Full	Reranker	SABR	Δ_{sel}	95% interval
Qwen2.5-7B	74.56	70.16	73.34	**+3.18**	[−0.06, 6.50]
Qwen2.5-14B	81.58	74.61	78.81	**+4.19**	[1.25, 7.19]
Mistral-7B	50.39	47.49	51.72	**+4.23**	[1.72, 6.76]

5.2 The aggregate boundary

The benefit is not uniform across tasks or budgets. At 70%, Mistral HotpotQA is −2.47 [−4.93, 0.01] versus reranking. The primary four-task, three-generator macro intervals cross zero against reranking at both caps (Table 3). At 50%, compression loses 3.17 points relative to full context. At 70%, the full-context difference is near zero, but its interval does not establish non-inferiority at a one-point tolerance. Appendix B retains all task–generator comparisons.

Table 3: Primary macro differences in paired task-score points. The macro equally weights four tasks and three generators while preserving within-question dependence.

Context cap	SABR minus reranker	SABR minus full context
50%	+0.24 [−1.01, 1.46]	−3.17 [−4.35, −1.99]
70%	+0.92 [−0.08, 1.91]	−0.17 [−1.08, 0.76]

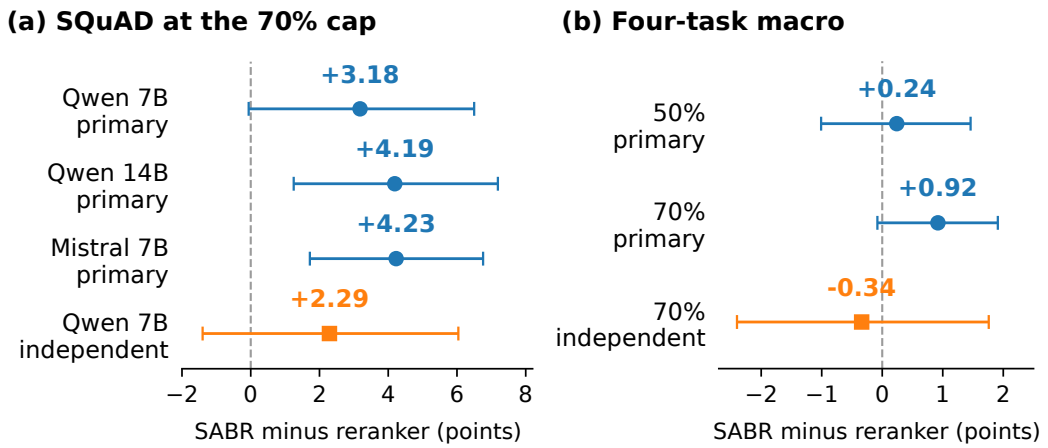

Figure 2: Localized benefit and aggregate uncertainty. Left: SQuAD at 70%, primary ($n = 346$ per generator) and independent Qwen-7B ($n = 256$). Right: primary and independent four-task macro contrasts. Circles denote primary estimates; squares denote independent estimates. Bars are paired pointwise 95% intervals; cohorts are not pooled.

5.3 Independent confirmation and component interventions

On the separate 896-question cohort, the SQuAD estimate remains positive, while the macro selection estimate is slightly negative (Table 4). Both intervals include zero. Thus the SQuAD direction persists, but an aggregate benefit is not independently confirmed. The negative macro full-context point estimate further separates selection value from quality preservation; the corresponding two-comparison Holm-adjusted test does not reject at 5% (Appendix C).

Table 4: Independent Qwen2.5-7B confirmation at 70%. SQuAD has 256 questions; macro comparisons use all 896.

Scope	Contrast	Difference [95% interval]
SQuAD	SABR minus reranker	$+2.29$ $[-1.40,\ 6.04]$
Four-task macro	SABR minus reranker	-0.34 $[-2.40,\ 1.76]$
Four-task macro	SABR minus full context	-2.05 $[-4.02,\ -0.11]$

Removing semantic scoring, entity bridging, or the boundary prior changes final inputs for 88.17%, 30.47%, and 9.71% of confirmation questions, respectively. Complete SABR minus those removals gives macro differences of -0.17, -0.25, and -0.08 points; all intervals include zero. These controls do not identify a globally beneficial individual factor. Their unequal input-change rates also caution against treating them as equally strong interventions.

5.4 Sentence units improve measured evidence retention

The sentence-unit pilot examines whether retaining coherent units changes the evidence available to the generator. Frozen semantic scores select whole sentences, while matched reranker and SABR inputs retain exactly the same realized context and prompt lengths. This is a separate representation change, not an enlargement of the primary reserve.

Table 5: Sentence-unit pilot ($n = 64$ per task). SQuAD retention is normalized answer-string presence; HotpotQA retention is mean within-question intact-support-sentence fraction, *not* the fraction of questions retaining all evidence.

Task	Method	Answer F1	Retention (%)	Context tokens
SQuAD	Sentence units	82.87	95.31	110.69
	Matched reranker	75.01	84.38	110.69
	Matched SABR	78.49	93.75	110.69
	Full context	88.53	—	182.20
HotpotQA	Sentence units	57.37	96.88	1677.92
	Matched reranker	55.62	83.07	1677.92
	Matched SABR	57.68	82.14	1677.92
	Full context	61.44	—	2410.63

Relative to reranking, sentence units increase answer-string presence by 10.94 percentage points on SQuAD and mean intact-support fraction by 13.80 points on HotpotQA. The two-task answer-F1 difference is $+4.80$ $[-0.28,\ 10.16]$; versus matched SABR it is $+2.03$ $[-2.30,\ 6.78]$. Retention improves, but neither answer-quality interval establishes a gain.

Whole units can leave unused space: mean context retention is 59.05% on SQuAD and 69.58% on HotpotQA. Two SQuAD contexts that cannot fit one whole sentence use a fixed prefix fallback and remain in every estimate. These properties prevent treating the pilot as another exact-70%-retention primary result.

6 Diagnostics: Structure, Representation, and Failure

6.1 A matched-prefix no-structure control

A separate Mistral diagnostic compares SABR with a no-structure control on a separate 1,498-question reserve. The control preserves SABR's semantic ensemble and initial lexical, entity-like, title, and boundary relevance scores, but removes adaptive bridging, neighbor propagation, redundancy updates, and sentence-group mechanisms. After selection, a symmetric prefix adapter clips both inputs to the smaller realized length and reconciles prompt-token boundaries. All 1,498 pairs match both context and prompt counts.

This control addresses a narrower question than SABR versus reranker: whether the coordinated policy changes answer quality when initial relevance is retained. However, adaptive versus ranked packing and the prefix adapter also change the evaluated policy. The diagnostic is not untouched confirmation and does not prove a unique causal mechanism. Cached helper scores are used here; recorded generation time is not end-to-end latency.

Table 6: Separate matched-prefix diagnostic: SABR minus its no-structure control. BoolQ uses parsed accuracy; other tasks use legacy token F1. These are *not* the audited primary benchmark scores. The mixed aggregate is exploratory.

Task	n	Metric	SABR	Control	Difference [95% interval]
BoolQ	384	Accuracy	74.22	72.14	$+2.08$ $[-1.04, 5.21]$
DROP-span	384	Legacy F1	24.12	21.97	$+2.14$ $[-0.39, 4.73]$
HotpotQA	346	Legacy F1	58.86	57.18	$+1.68$ $[-0.34, 3.77]$
SQuAD	384	Legacy F1	53.43	50.76	$+2.67$ $[-0.12, 5.45]$
Equal-task mean	1,498	Mixed	52.66	50.51	$+2.15$ $[0.81, 3.50]$

All task point estimates are positive, but each task interval crosses zero. The equal-task mixed diagnostic is positive. Because it combines parsed accuracy with legacy F1 on a previously evaluated reserve, with a post-hoc scoring correction and no multiplicity adjustment, it is secondary evidence about this control, not a replacement for the uncertain primary or confirmation macro result. BoolQ parser sensitivity and the exact adapter are retained in Appendix D.

6.2 Evidence presence is not answer sufficiency

Evidence retention and answer quality are different endpoints. Among sentence-pilot matched-SABR answers that are worse than full context, 11 of 15 SQuAD cases retain an answer string, and 6 of 7 HotpotQA cases retain the measured supporting evidence (Table 7). These observations show why a retention metric cannot replace answer evaluation.

Table 7: Matched SABR versus full context in the sentence pilot. Each task has 64 questions. Retained/missing evidence partitions worse-answer cases using the task-specific diagnostic; these are not causal error labels.

Task	Worse answers	Better answers	Retained evidence	Missing evidence
SQuAD	15	6	11	4
HotpotQA	7	5	6	1

An answer string can survive while its qualifier is lost; a retained supporting sentence can be ignored or misinterpreted. Conversely, removing distracting text can improve some answers. The saved diagnostics do not determine which explanation caused each outcome. We therefore report observable evidence and answer degradation separately, rather than assigning every failure to either the selector or the generator.

7 Runtime: Shorter Inputs Are Not Automatically Faster

7.1 Native-length timing with fixed output length

The original Qwen diagnostic measures semantic scoring, selection, rendering, tokenization, generation, and decoding on twelve contexts with resident models and fixed 32-token outputs. Five arms and two measured repetitions give 120 observations; twenty warmups are excluded. Cold-start, network, queueing, and segmentation costs are outside the measurement. The independent unit remains the context.

Table 8: Native-length Qwen timing ($n = 12$). A full/method ratio above one indicates a speedup over full context. Ratios are medians of paired context ratios, *not* ratios of displayed median seconds.

Method	Cap	Seconds	Full/method ratio	95% interval
Full context	100%	1.159	1.000	[1.000, 1.000]
Reranker	50%	1.377	0.842	[0.809, 0.868]
Reranker	70%	1.499	0.774	[0.735, 0.797]
SABR	50%	1.194	0.992	[0.966, 1.002]
SABR	70%	1.268	0.939	[0.909, 0.944]

SABR does not demonstrate a full-context speedup in this setup. It is faster than the implemented reranker, but different packing and repeated tokenization costs affect that comparison. This is an implementation result, not a complexity claim that two-helper selection is intrinsically cheaper than reranking.

7.2 Exact optimization and longer inputs

An A40/Mistral engineering retest replaces repeated redundancy and nearest-distance calculations with exact incremental updates. Selection matches the original on 1,498 examples and 90 edge tests, and 96 fresh paired generations match. On 32 contexts with three repetitions, optimized mean latency is 0.673 s versus 0.826 s for original SABR: 18.55% lower time. It remains 12.14% slower than full context (Table 9). Timed requests recompute helper scores; equivalence checks do not substitute cached scoring for online work.

Table 9: Mistral timing extensions. Here ratios are *method/reference means*, so values below one indicate lower latency. These contexts, estimators, and output settings differ from Table 8.

Contexts	Contrast	Time ratio [95% interval]
32	Incremental/original SABR	0.814 [0.762, 0.869]
32	Incremental SABR/full context	1.121 [1.013, 1.246]
38	Incremental SABR/full context	1.068 [1.021, 1.114]
38	Incremental SABR/reranker	0.766 [0.710, 0.824]

The long-input challenge uses 38 train-derived HotpotQA prompts spanning 4,105–6,986 full-prompt tokens, with three repetitions per method. Full context, incremental SABR, and reranker take 1.250, 1.335, and 1.742 s on average. SABR's legacy-F1 difference from full context is $+2.41$ $[-1.11, 7.72]$, not an established quality gain. The compressed arms share a 70% cap but *not* identical realized lengths. All 342 successful requests are included; the failed precursor session is excluded as a whole. These tests quantify engineering progress without establishing a quality–latency frontier, production throughput advantage, or break-even input length (Appendix E).

8 Discussion and Limitations

Where the evidence supports coordinated selection. The primary SQuAD result answers the paper's controlled question: at the same realized input length, SABR's complete selection policy improves the point estimate by 3–4 F1 across three generators, with two paired intervals excluding zero. This rules out retaining more tokens as an explanation. The separate confirmation cohort keeps the SQuAD effect positive but uncertain, while both primary and confirmation macro results limit a general cross-task claim. The appropriate conclusion is a localized selection benefit, not universal superiority.

What "structure-aware" does and does not identify. The primary SABR–reranker contrast changes semantic mixing and coordination together. The matched-prefix no-structure diagnostic preserves initial relevance and gives additional evidence for the coordinated policy, but uses a different reserve, adapter, and scoring protocol. Independent single-factor removals do not establish a globally beneficial component. Together, these analyses motivate coordinated retention without claiming that adjacency, entity bridging, or any individual heuristic has been causally established as the source of every gain.

Representation is part of the method. Eight-word fragments and retained title information differ from common paragraph-based retrieval. Some benefit may come from repairing relations broken by this representation. Sentence-level selection raises measured evidence retention, but its answer-quality intervals remain uncertain and it sometimes leaves budget unused. The result is useful precisely because it separates a change in what survives from a guaranteed change in what the generator answers. Generalization to other segmentation schemes, richer document structures, and substantially longer inputs remains open.

Development, metrics, and uncertainty. The primary and matched-prefix reserves were previously inspected. Public-validation confirmation reduces reuse within this research workflow but cannot rule out model-pretraining exposure. Paired resampling respects model dependence within questions; it does not remove development bias or automatically handle every form of source-document dependence. Task intervals are pointwise, while declared aggregate families use multiplicity adjustment. Restricted DROP coverage and answer-only HotpotQA scoring limit benchmark claims. Legacy-scored extensions are explicitly separated from the audited primary metrics.

Engineering improvements are not serving guarantees. The incremental implementation preserves tested selections and reduces overhead, yet its average runtime remains above full context in both the 32-context retest and 38-context longer-input challenge. Warm, batch-one measurements do not characterize cold starts, concurrent serving, or production traffic. Different retained lengths and naturally varying output lengths in the longer-input test prevent interpreting its timing contrast as an exact-token causal comparison. Broader scaling or scoring-amortization benefits are not credited without measurement.

Auxiliary policies are not the core contribution. The optional learned escalation pilot does not establish an allocation advantage over matched-quota controls. Its sequential cost must include both the compressed and full-context generations on escalated questions. A separate legacy answer-correctness calibration diagnostic does not calibrate the primary deterministic selector. These auxiliary extensions are retained in the appendices rather than used to strengthen the main claim.

9 Conclusion

SABR makes *what should survive together* an explicit decision in fixed-budget evidence selection. Its frozen semantic scores, local coordination, and exact token packing produce a concrete training-free alternative to independent relevance ranking. The strongest evidence is a 3–4-point primary SQuAD improvement at a 70% cap across three generators; independent testing and complete task comparisons define its limits. Representation and matched-prefix diagnostics add evidence about retention, while timing studies separate implementation savings from end-to-end speedups. The broader lesson is to compare selectors at the same realized input length and evaluate retained evidence, answer quality, and online cost as distinct outcomes.

AI USE STATEMENT

Generative AI tools were used for literature retrieval, language polishing, assistance with experimental code development and revision.

ETHICS STATEMENT

The study uses public question-answering datasets and released model checkpoints under their respective terms. Context deletion can remove qualifications or supporting facts while leaving a fluent answer, so the reported benchmark scores and evidence-retention diagnostics do not constitute reliability guarantees for high-stakes use. No conclusion about protected groups, clinical decision making, or safety-critical deployment follows from these aggregate results. Reference answers and supporting-fact labels are used only for evaluation and diagnostics, not by the selection policy.

REPRODUCIBILITY STATEMENT

Appendix A specifies the evaluated selector, helper configurations, rendering assumptions, and scoring protocol. Appendices B–E retain complete comparisons and distinguish cohorts, matching rules, inference units, and measurement boundaries. The manuscript source bundle includes editable tables, the plotted numerical values, selected machine-readable extension summaries, and a provenance ledger. These materials document the reported evidence but are not a replacement for the full prediction and runner archives. Large model weights are not bundled. Historical and failed sessions are not silently merged into successful evaluations.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023. https://arxiv.org/abs/2310.11511.

Yushi Bai et al. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of ACL*, 2024. https://aclanthology.org/2024.acl-long.172/.

Beijing Academy of Artificial Intelligence. *BAAI/bge-reranker-base model card*, 2023a. https://huggingface.co/BAAI/bge-reranker-base.

Beijing Academy of Artificial Intelligence. *BAAI/bge-small-en-v1.5 model card*, 2023b. https://huggingface.co/BAAI/bge-small-en-v1.5.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL-HLT*, pp. 2924–2936, 2019. https://doi.org/10.18653/v1/N19-1300.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of NAACL-HLT*, pp. 2368–2378, 2019. https://doi.org/10.18653/v1/N19-1246.

Taeho Hwang, Sukmin Cho, Soyeong Jeong, Hoyun Song, SeungYoon Han, and Jong C. Park. EXIT: Context-aware extractive compression for enhancing retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL*, pp. 4895–4924, 2025. https://doi.org/10.18653/v1/2025.findings-acl.253.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of NAACL-HLT*, pp. 7036–7050, 2024. https://doi.org/10.18653/v1/2024.naacl-long.389.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of EMNLP*, pp. 13358–13376, 2023. https://doi.org/10.18653/v1/2023.emnlp-main.825.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of ACL*, pp. 1658–1677, 2024. https://doi.org/10.18653/v1/2024.acl-long.91.

Vladimir Karpukhin, Barlas Oğuz, 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, 2020. https://aclanthology.org/2020.emnlp-main.550/.

Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020. https://arxiv.org/abs/2005.11401.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. https://aclanthology.org/2024.tacl-1.9/.

Mistral AI. *Mistral-7B-Instruct-v0.3 model card*, 2024. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL*, pp. 963–981, 2024. https://doi.org/10.18653/v1/2024.findings-acl.57.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. https://arxiv.org/abs/2412.15115.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP*, pp. 2383–2392, 2016. https://doi.org/10.18653/v1/D16-1264.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pp. 3982–3992, 2019. https://aclanthology.org/D19-1410/.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with compression and selective augmentation. In *International Conference on Learning Representations*, 2024. 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. https://doi.org/10.18653/v1/D18-1259.

A Frozen Selector and Scoring Details

A.1 Helpers, routes, and initial scores

The bi-encoder is `BAAI/bge-small-en-v1.5`, with 192-token inputs, normalized mean pooling, and batches of 96. The cross-encoder is `BAAI/bge-reranker-base`, with question–fragment pairs truncated to 256 tokens and batches of 64. Both operate in evaluation mode. Within-context min–max normalization maps a constant score vector to 0.5. Multi-document detection precedes boolean and arithmetic patterns. Boolean detection uses auxiliary-verb openings; arithmetic detection uses fixed quantity/comparison phrases; other inputs use the factoid route. These routes are deterministic text rules, not dataset labels or the learned escalation model.

Let L_i be query-normalized inverse-document-frequency term overlap, E_i overlap with deterministic entity-like spans, T_i title overlap, and $B_i = \mathbf{1}\{i \in \{1, m\}\}$ the boundary indicator. Entity-like spans are capitalized, underscored, or numerical patterns, not trained semantic entity labels. With s_i from Equation 3, the initial relevance score is

$$b_i = 2.2L_i + 1.35E_i + 0.9T_i + w_r s_i + 0.04B_i, \tag{4}$$

where $w_r = 2$ for multi-document input and 0.75 otherwise. Adaptive packing uses the incremental score

$$g(i \mid S) = b_i + 0.05E(i, S) + 0.02R(i, S) + P(i, S) - 0.08J(i, S). \tag{5}$$

Entity overlap E is capped at two; rare-term overlap R is capped at three. Proximity P contributes 0.12 for adjacent fragments and 0.04 at distance two. J penalizes redundancy with selected fragments. Ties favor earlier source fragments. For the non-boolean/non-multi-document route with numeric density below 25%, semantic relevance is propagated within three positions before adaptive packing.

The high-numeric-density branch instead forms consecutive sentence groups G and scores them by

$$z_G = \max_{i \in G} s_i + 0.5 \frac{|Q \cap T(G)|}{\max(1, |Q|)} + 0.1(e_G + n_G), \tag{6}$$

where Q is the set of query terms, $T(G)$ the group's terms, and the entity/number counts e_G, n_G are each capped at three. Whole groups are packed before individually ranked fill. All coefficients are frozen for the reported evaluations; their design was informed by development data.

A.2 Budget adapter and what it controls

The internal budget is a packing aid, not the reported tokenizer budget. After every proposal, selected fragments are rendered in source order and counted with the generator tokenizer. Over-budget proposals trigger a smaller internal budget and another attempt, up to twenty times; direct-token pruning removes non-fitting fragments if needed. Returned inputs satisfy Equation 1. The primary matched control additionally checks both equalities in Equation 2. Passing the same nominal cap alone is not an exact-length control, as illustrated by the separate longer-input test.

A.3 Scoring and integrity checks

The primary answer extractor takes the final `FINAL:` segment when present, otherwise the last nonempty line. Extraction is gold-blind. SQuAD and HotpotQA then use task-appropriate answer normalization and available aliases. BoolQ uses strict yes/no accuracy. DROP is restricted to the single-span protocol. Full-context answers use the same supplied segmented representation as compressed arms, not independently retrieved or differently formatted context.

The primary archive contains 20,760 quality records: 1,384 questions, three generators, and five full/compressed conditions. Token, prompt, key, and target checks found no mismatches in the paired SABR/reranker arms. Recomputing scoring and paired analyses from cached answers recovers the reported primary tables without new generation. Recorded software includes PyTorch 2.8.0+cu128, Transformers 4.56.2, Tokenizers 0.22.2, and NumPy 2.1.2. This records the environment used; it is not a numerical-equivalence claim for later versions.

B Complete Audited Primary Comparisons

Every task below has 346 shared questions per generator. Positive differences favor SABR. Confidence intervals are paired, pointwise 95%; they are not adjusted across the individual task–generator cells. The macro analysis resamples questions within task while preserving all three model outcomes for a sampled question. It does not treat a model–question record as an independent question.

Table 10: Complete primary comparison at the 70% cap. Differences use audited task-specific answer scores on a 0–100 scale.

Generator	Task	Δ_R	95% interval	Δ_F	95% interval
Qwen2.5-7B	BoolQ	$+0.87$	$[-1.73, 3.47]$	-1.45	$[-4.05, 1.16]$
	SQuAD	$+3.18$	$[-0.06, 6.50]$	-1.22	$[-3.64, 1.19]$
	HotpotQA	-0.27	$[-2.59, 2.03]$	-0.02	$[-2.65, 2.60]$
	DROP	$+2.48$	$[-1.79, 6.71]$	$+0.36$	$[-3.55, 4.30]$
Qwen2.5-14B	BoolQ	-1.16	$[-3.76, 1.16]$	$+1.45$	$[-1.45, 4.34]$
	SQuAD	$+4.19$	$[1.25, 7.19]$	-2.78	$[-5.09, -0.61]$
	HotpotQA	-0.37	$[-3.23, 2.51]$	-2.26	$[-5.07, 0.43]$
	DROP	$+0.50$	$[-3.62, 4.58]$	$+0.36$	$[-3.23, 3.90]$
Mistral-7B	BoolQ	0.00	$[-3.18, 2.89]$	$+1.73$	$[-1.73, 5.20]$
	SQuAD	$+4.23$	$[1.72, 6.76]$	$+1.34$	$[-1.19, 3.92]$
	HotpotQA	-2.47	$[-4.93, 0.01]$	-0.24	$[-2.60, 2.17]$
	DROP	-0.09	$[-2.77, 2.68]$	$+0.75$	$[-1.61, 3.18]$

Table 11: Complete primary comparison at the 50% cap. Δ_R denotes SABR minus reranker; Δ_F denotes SABR minus full context.

Generator	Task	Δ_R	95% interval	Δ_F	95% interval
Qwen2.5-7B	BoolQ	-0.58	$[-3.76, 2.60]$	-6.36	$[-9.83, -2.89]$
	SQuAD	$+1.38$	$[-3.01, 5.62]$	-6.76	$[-10.24, -3.21]$
	HotpotQA	-1.91	$[-4.76, 0.99]$	-1.33	$[-4.80, 2.04]$
	DROP	$+2.38$	$[-2.30, 7.03]$	-3.53	$[-7.90, 0.77]$
Qwen2.5-14B	BoolQ	$+0.58$	$[-2.31, 3.47]$	-2.60	$[-6.07, 0.58]$
	SQuAD	$+0.91$	$[-2.92, 4.61]$	-8.25	$[-11.18, -5.40]$
	HotpotQA	-0.50	$[-3.17, 2.18]$	-1.23	$[-4.52, 2.05]$
	DROP	$+1.62$	$[-2.80, 5.98]$	-3.04	$[-7.41, 1.28]$
Mistral-7B	BoolQ	-1.16	$[-4.62, 2.31]$	-2.31	$[-6.65, 2.02]$
	SQuAD	$+1.28$	$[-2.09, 4.69]$	-1.98	$[-5.06, 1.10]$
	HotpotQA	-2.04	$[-4.80, 0.67]$	-1.89	$[-4.53, 0.76]$
	DROP	$+0.86$	$[-2.09, 3.83]$	$+1.27$	$[-1.20, 3.75]$

Across the four primary macro contrasts (two caps, each versus reranker and full context), Holm-adjusted sign-flip p values are 1.000 for 50% versus reranker, 0.0004 for 50% versus full, 0.200 for 70% versus reranker, and 1.000 for 70% versus full. An interval crossing zero is not an equivalence result. The three positive SQuAD point estimates also do not replace an independent cohort test.

C Independent Confirmation and Sentence-Pilot Details

The 896-question public-validation cohort is fixed before generation, with no favorable seed selection or outcome-based stopping in this confirmation. All six arms are evaluated on the same questions. Reusing an answer for identical final input saves computation but provides no additional independent evidence. Task-level comparisons are exploratory; the two macro comparisons constitute the primary confirmation family.

Table 12: Independent task differences at 70%: SABR minus comparator. Task intervals are paired and pointwise.

Task	Versus reranker	Versus full context
BoolQ	-1.17 $[-4.30, 1.95]$	-2.34 $[-5.47, 0.78]$
SQuAD	$+2.29$ $[-1.40, 6.04]$	-0.47 $[-3.39, 2.51]$
HotpotQA	-0.11 $[-3.29, 3.01]$	-1.32 $[-5.07, 2.41]$
DROP-span	-2.37 $[-8.33, 3.52]$	-4.05 $[-9.46, 1.28]$

The macro full-context difference is -2.05 $[-4.02, -0.11]$. Its pointwise interval excludes zero, but its two-comparison Holm-adjusted sign-flip p value is 0.092. Accordingly, the main text distinguishes a negative point estimate from an adjusted significance claim.

Table 13: Independent single-factor removals. Positive differences would favor complete SABR. The macro component family has three comparisons.

Removed factor	Changed inputs (%)	SABR minus removal [95% interval]
Semantic scoring	88.17	-0.17 $[-2.11, 1.76]$
Entity bridging	30.47	-0.25 $[-0.93, 0.43]$
Boundary prior	9.71	-0.08 $[-0.33, 0.11]$

All three Holm-adjusted component p values are 1.000. The localized SQuAD semantic-removal contrast is positive before the twenty-cell adjustment but not after it. Component effects are reported from outcomes and input changes rather than inferred from heuristic names.

C.1 Representation and evidence diagnostics

The sentence-unit pilot is separate from both the primary reserve and confirmation cohort. Its matched methods have the same actual context and prompt counts for each question, but whole-sentence packing need not fill the nominal cap. Two SQuAD prefix fallbacks remain in the reported sample. Evidence-retention screening and the answer-quality cohort are distinct stages, and their counts are not added together.

SQuAD retention measures normalized answer-string presence. HotpotQA retention averages the fraction of intact gold supporting sentences within each question. The latter is not the percentage of questions retaining every supporting fact, and neither diagnostic proves preservation of every necessary qualifier. Full-context evidence percentages are not reported in the saved pilot; no unmeasured percentages are filled into Table 5.

Among sentence-selection answers worse than full context, 9 of 12 SQuAD cases retain an answer string and 8 of 9 HotpotQA cases retain the measured supporting evidence. For matched reranker, the corresponding counts are 9 of 19 and 7 of 11. Matched-SABR counts appear in Table 7. These are descriptive partitions of worse-answer cases, not mutually exclusive causal explanations of all errors. The independent cohort has answer-level comparisons but no equivalent saved evidence annotations; the sentence-pilot labels are not transferred to it.

D Matched-Prefix Structure Diagnostic

D.1 Frozen configuration and symmetric length adapter

The diagnostic uses Mistral-7B-Instruct-v0.3 on an A40 with a 70% context cap, BF16, SDPA, greedy generation, batch size eight, and at most 64 new tokens. The recorded reserve has 384 BoolQ, 384 DROP-span, 346 HotpotQA, and 384 SQuAD questions. Seed buckets designate sample partitions, not independent model-training replications. Cached semantic scores are reused for selection; this protocol supplies answer-quality contrasts, not a complete online latency measurement.

The no-structure arm preserves the semantic mixture and the initial lexical, query–entity, title, and boundary terms. It removes adaptive bridging, local propagation, redundancy updates, and sentence-group mechanisms, then packs ranked fragments. A symmetric prefix adapter clips both contexts to the smaller realized length and removes at most twelve additional tokens to reconcile prompt-token boundaries. The completed run contains 2,996 predictions forming 1,498 matched pairs, with zero audited context/prompt-token mismatches.

This comparison controls initial relevance more closely than the primary reranker contrast. Nevertheless, the adapter changes final retained text, and the ranked control does not preserve every feature of adaptive packing. The result is evidence about these frozen policies under the recorded adapter, not an isolated causal estimate of a single structural feature. The reserve is not labeled an untouched confirmation set.

D.2 BoolQ parsing and mixed-score interpretation

The fixed primary parser for this diagnostic, `final_or_leading`, uses a recognized final answer or falls back to a leading answer. The reported alternatives are sensitivity analyses, not candidates from which to choose the largest gain. Invalid/unparsed outputs are retained and counted as incorrect.

Table 14: BoolQ parser sensitivity on the same 384 matched-prefix pairs. All values are accuracy percentages or percentage-point differences.

Parser	SABR	Control	Difference [95% interval]
Final or leading (reported)	74.22	72.14	$+2.08$ $[-1.04, 5.21]$
Leading only	79.17	76.82	$+2.34$ $[-0.52, 5.21]$
Final only	7.29	7.29	0.00 $[-2.60, 2.60]$

Under the reported parser, SABR helps 23 questions and harms 15 relative to its control; 60 SABR and 63 control outputs are unparsed. Leading-only parsing has 36 and 38 unparsed outputs, while final-only parsing has 351 and 348. This sensitivity demonstrates why an answer-format convention must be specified instead of equating generic token overlap with yes/no accuracy.

Table 6 combines corrected BoolQ accuracy with the unchanged legacy token-F1 scores for the other three tasks. It is therefore labeled a mixed diagnostic rather than official benchmark macro F1. The equal-task aggregate is $+2.15$ $[0.81, 3.50]$, whereas all four task-level intervals include zero. This exploratory, post-hoc scoring analysis is not multiplicity-adjusted and does not remove method-development uncertainty. Its sample size and interval are not combined with the primary $+0.92$ macro contrast, the independent -0.34 contrast, or the distinct learned-compressor experiment.

E Timing Scope, Exact Optimization, and Longer Inputs

E.1 Original native-length diagnostic

The twelve-context Qwen study contains five methods and two timed repetitions after warmup: 120 measured observations and twenty excluded warmups. Each of four length bins contains three contexts; the nominal longest bin does not cover its entire upper range. Context bootstrap intervals pair methods and resample contexts. Repetitions do not multiply the independent sample size. Fixed 32-token generation makes output length common in this diagnostic.

All helper scoring and selection in the measured path are included, along with rendering, tokenization, generation, and decoding. Cold-start, network, queueing, and segmentation are excluded. Table 8 reports median seconds and the median of paired full/method context ratios, so taking the ratio of displayed medians does not generally reproduce its ratio column. The paired SABR speed ratios relative to the implemented reranker are 1.162 and 1.184 at 50% and 70%. Packing implementation and repeated tokenization affect these comparisons.

E.2 Mistral engineering measurements and exactness checks

The separate 32-context Mistral set contains eight questions per task, selected by input-length strata from a train-derived reserve. Original full contexts span 68–3,471 context tokens. The first engineering measurement gives original SABR mean 0.831 s and full context 0.601 s, with time ratio 1.383 $[1.212,\ 1.575]$. The later exact-incremental retest remeasures both arms; it does not mix their earlier times with new optimized measurements.

The incremental implementation updates redundancy and nearest-selected distance instead of recomputing the corresponding quantities. It is an engineering change to the same tested selections, not a new quality algorithm. The archive records equality on all 1,498 selection examples and ninety edge cases, plus identical outputs in ninety-six fresh original/optimized generation pairs. These checks establish equality on the tested cases, not a proof for every possible input. Cached scores are used only in CPU equivalence checks; every timed GPU request recomputes the helper scores.

Table 15: Exact-incremental Mistral retest: 32 contexts, three repetitions per arm (288 timed requests). Time ratios in Table 9 bootstrap context-level mean times.

Implementation	Mean seconds
Full context	0.600227
Original SABR	0.826391
Incremental SABR	0.673065

The paired bootstrap is stratified by task after averaging repetitions within context. The optimized/original ratio is 0.814464 $[0.762053,\ 0.869197]$; optimized/full is 1.121352 $[1.013197,\ 1.246246]$. The 18.55% reduction therefore concerns original-selector overhead, not a full-context speedup. These warm A40, batch-one measurements exclude cold-start, network, queueing, and disk logging. Output lengths may vary across different methods. They should not be equated with the fixed-output Qwen study or interpreted as production throughput.

A previously reported request-local CPU microbenchmark preserved 288 tested output hashes while saving approximately 8–11 ms of selector time. That older microbenchmark is distinct from the completed end-to-end incremental retest above; it is no longer the only available optimization evidence. Neither engineering result is paired with the different matched-prefix quality protocol to construct a synthetic quality–latency frontier.

E.3 38-CONTEXT LONGER-INPUT CHALLENGE

The longer-input set contains the 38 HotpotQA contexts removed by the prior 4,096-token prompt limit from a 1,536-example candidate reserve, leaving 1,498 shorter examples. The archive checks identifier, group, and normalized-question disjointness from the shorter reserve. Selection is based on input length, not favorable answer outcomes. The recorded full prompts span 4,105–6,986 tokens; raising the maximum input limit to 8,192 does not imply measurements throughout the entire 8k range.

The successful V2 run uses an A40, Mistral-7B-Instruct-v0.3, BF16, batch-one greedy generation, and at most 64 new tokens. Full context, reranker, and incremental SABR are measured three times on each context: 342 unique requests, 114 per method. All 114 context–method combinations produce identical answers across their repetitions; answer-quality scoring uses repetition zero only. One reranker context reaches the 64-token generation cap and remains included. A failed precursor session is preserved but excluded in its entirety, rather than selectively merged into V2.

Table 16: Longer-input challenge, 38 contexts. Seconds are averages over context-level repetition means. Answer F1 is the recorded legacy metric; exact match is shown as a count of 38.

Endpoint	Full context	Reranker	Incremental SABR
Total time (s)	1.249722	1.741833	1.334665
Helper scoring (s)	0.000009	0.131039	0.174393
Selection (s)	0.006648	0.617522	0.217114
Generation pipeline (s)	1.243065	0.993272	0.943158
Context tokens	4689.63	3279.63	3212.39
Output tokens	13.18	13.55	12.45
Peak allocated memory (GiB)	15.78	15.45	15.44
Legacy answer F1	59.54	60.18	61.96
Exact match	18/38	19/38	19/38

Table 17: Paired longer-input contrasts, using 5,000 context bootstraps. Time ratios are ratios of mean times; quality differences are legacy F1 points. Intervals are exploratory and unadjusted.

Contrast	Time ratio [95% interval]	F1 difference [95% interval]
SABR/full	$1.068\ [1.021,\ 1.114]$	$+2.41\ [-1.11,\ 7.72]$
SABR/reranker	$0.766\ [0.710,\ 0.824]$	$+1.78\ [-4.72,\ 8.92]$
Reranker/full	$1.394\ [1.297,\ 1.498]$	$+0.64\ [-8.18,\ 8.70]$

The bootstrap uses seed 26092411 and resamples contexts after averaging repetitions. Mean context reduction for SABR is approximately 31.50%. Its generation pipeline saves 0.300 s relative to full context, while scoring and selection add approximately 0.385 s, leaving a net 0.085 s increase. Twelve of 38 contexts are individually faster under SABR, but the aggregate remains slower. This decomposition is measured accounting, not a causal decomposition with all other factors held fixed.

Reranker and SABR satisfy the same 70% context cap but retain different realized counts; output lengths also differ. Their timing comparison therefore does not have the exact-length control of the main quality experiment. The train-derived, length-selected cohort is an exploratory challenge, not an official long-context benchmark or production traffic sample. No statistically established quality gain, break-even length, or production-throughput benefit follows from these data.

F Separate Historical Comparisons and SABR-L Extension

F.1 Token-cap-adapted learned compression

An earlier Mistral protocol uses a different 1,498-context reserve and a custom token-F1 metric. At 70%, its SQuAD means are 53.65 for full context, 48.34 for a token-cap-adapted LLMLingua-2, 49.64 for reranker, and 53.10 for SABR. SABR minus the adapted learned compressor is $+4.76$ $[1.72, 7.84]$; versus reranker it is $+3.46$ $[1.09, 5.82]$. On HotpotQA the corresponding contrasts are $+6.01$ $[3.05, 9.03]$ and -0.33 $[-2.09, 1.49]$.

These values support comparison with that specific adapted baseline on its own cohort. They are not substituted for the audited primary Mistral SQuAD estimate of $+4.23$, pooled with the matched-prefix structure diagnostic, or presented as superiority over all learned compression. Historical LongBench subsets and 32B runs also use different protocols; their non-positive comparisons limit extrapolation beyond the supplied-context setting.

F.2 Frozen lexical/structural Mistral configuration

The later V18 runner uses Mistral-7B-Instruct-v0.3 on an A40 with seeds 1201–1203 and estimated context ratios 0.5, 0.7, and 1.0. All phases completed, with final `COMPLETE` at 22:43:34 CDT on September 22, 2026. Its 84 summary keys are unique. The recorded command does not enable the semantic selector; the inherited lexical/structural path is therefore denoted *SABR-L*, not the semantic primary SABR method. Controls are full context, head–tail, and lexical top-k.

The later paired export joins model, experiment, seed, and example ID, then averages per-record SABR-L-minus-lexical legacy-F1 differences. It uses 5,000 within-task paired-row bootstrap replicates. This is not an exactly token-matched reranker comparison, a source-question-clustered confirmation, or additional independent evidence for the primary sample.

Table 18: Separate estimated-70% SABR-L run: paired legacy token-F1 differences from lexical top-k. These exploratory row-bootstrap intervals are not multiplicity-adjusted. BoolQ is legacy F1 here, not accuracy.

Task	Difference	Reported 95% interval
Single-hop position	-12.85	$[-17.01, -9.03]$
SQuAD	$+1.72$	$[-2.63, 6.07]$
Two-hop composition	$+1.17$	$[-1.35, 3.70]$
Conflict/decoy	0.00	[0.00, 0.00]
HotpotQA	$+3.03$	$[-0.70, 6.82]$
DROP subset	$+0.28$	$[-4.11, 4.58]$
BoolQ (legacy F1)	$+1.22$	$[-1.14, 3.86]$

The single-hop result identifies a failure of this configuration under the reported analysis. Positive task intervals cross zero. The conflict/decoy tie concerns the reported scores in this condition, not general equivalence. Earlier arm-summary DROP and BoolQ contrasts were -0.25 and $+0.92$, unlike the later paired export's $+0.28$ and $+1.22$; point estimates from one aggregation must not be combined with intervals from another. Stored IDs and seeds need not imply disjoint source questions, and the row bootstrap does not resolve that issue. These results remain supplementary rather than evidence of a significant multi-task gain.

G Auxiliary Escalation, Cost, and Calibration

G.1 Fixed-quota escalation audit

A learned wrapper was trained on an earlier 1,101-question cohort with a separate 283-question calibration cohort. The recorded threshold is 8.369600222374. Rescoring the previously inspected 256-question pilot does not refit its ranking, mapping, or threshold. All 2,560 quota decisions and 256 frozen-threshold decisions checked in that audit are unchanged. The deterministic input route in Equation 3 is not this learned wrapper.

Table 19: Learned routing minus comparator at a fixed 25% escalation quota: sixteen of sixty-four questions per task. All 256 questions use audited quality scores.

Comparator	50% cap	70% cap
Confidence	-0.20 $[-2.59, 2.23]$	$+0.13$ $[-1.90, 2.18]$
Fixed schedule	-0.15 $[-3.18, 2.91]$	$+1.54$ $[-0.58, 3.85]$
Random expectation	$+0.24$ $[-1.36, 1.88]$	$+0.01$ $[-1.30, 1.31]$

Fixed-hash random and uncertainty comparisons also have intervals crossing zero. Random expectation averages over a uniform exact-count mask, not a selected favorable random seed. The 50% beneficial-escalation AUC is 0.483. The frozen threshold escalates 61/256 questions at 50% and 62/256 at 70%. At 50%, audited base-error rates are 34.43% for escalated and 49.23% for non-escalated questions; beneficial-escalation rates are 11.48% and 13.33%. The selected group is not enriched for either event in this pilot. These descriptive rates are not risk guarantees.

Sequential deployment pays for both generations on escalated examples:

$$C_i = C_i^{\text{compressed}} + a_i C_i^{\text{full}}, \qquad a_i \in \{0, 1\}. \tag{7}$$

Generator-token costs include prompts and answers but exclude semantic helpers and packing. Relative to direct full context, the learned-policy values are:

Table 20: Sequential generator-token cost. These are not wall-clock speedups.

Escalation fraction	50% cap	70% cap
0%	0.547	0.729
25%	0.806	0.974
50%	1.056	1.236
100%	1.547	1.729

At a 25% quota, the 70% cap nearly exhausts direct-full generator-token cost even before helper costs. At full escalation, every example pays for two passes. The pilot does not establish a learned allocation advantage; low average calibration error alone would not demonstrate useful escalation ranking either.

G.2 Separate post-hoc answer-correctness calibration

A separate uncertainty-router score cache is joined to SABR-L answer records, not to the primary exact-token experiment. Both stored ID sets have 1,440 entries, with no missing or extra IDs. The join verifies IDs only; it does not establish equality of question text, block content, ordering, or preprocessing across runs.

The frozen `BAAI/bge-small-en-v1.5` score feature is the top-two gap after within-example min–max scaling:

$$x = \frac{s_{(1)} - s_{(2)}}{s_{\max} - s_{\min}}, \qquad s_{(1)} \geq s_{(2)}. \tag{8}$$

Constant score vectors return zero; fewer than two finite scores return one. This scalar is a relevance feature, not an answer probability. The binary outcome is the runner's normalized exact-match correctness field, not primary task-appropriate answer F1.

A deterministic hash of stored ID assigns 725 records to calibration and 715 to test. A one-dimensional logistic mapping is fitted on the calibration split, including feature centering and scaling. The baseline assigns every test record the calibration correctness rate. This post-hoc mapping neither changes selections nor re-runs generation.

Table 21: Auxiliary correctness calibration on 715 test records after fitting on 725 calibration records. Intervals use 2,000 stored-ID cluster bootstrap samples of the test split, holding the fitted mapping fixed.

Metric	Estimate	Reported 95% interval
ECE_{10} (equal-width bins)	0.0802	$[0.0573, 0.1162]$
Brier score	0.2239	$[0.2143, 0.2333]$
Negative log-likelihood	0.6345	$[0.6125, 0.6559]$
Constant-predictor Brier	0.2416	Not reported
Test exact-match correctness	0.5916	Not reported

The calibrated predictor's Brier point estimate is lower than the constant predictor's, but the recorded summary supplies no paired interval for the difference. These results do not calibrate the primary deterministic selector or establish an answer-quality advantage for escalation. Stored-ID disjointness has not been audited for duplicated source questions or passages across seeds, and the test bootstrap excludes calibration-fit uncertainty. The imported score cache is therefore auxiliary exploratory evidence, not independently audited reliability evidence.

G.3 Evidence boundaries

The results distinguish exact realized length from a common cap, audited task metrics from legacy or mixed diagnostics, independent question units from repeated measurements, and method changes from implementation changes. The separate extensions answer narrower questions than the primary matched-token experiment. They neither enlarge its sample size nor repair the absence of an independently confirmed aggregate gain. Retaining these boundaries makes it possible to report informative secondary findings without treating every positive point estimate as a replication of the same claim.