CONSIEVE: CONSTRAINT-SENSITIVE EVIDENCE SELECTION FOR RAG CONTEXT COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-retrieval context compressors for retrieval-augmented generation (RAG) typically select evidence based on its semantic relevance to the query. Relevance alone does not guarantee validity because text can remain topically relevant while violating an answer-critical condition, especially under temporal constraints, and therefore fail to support the answer. To measure this failure directly, we construct fixed-evidence counterfactual diagnostic pairs by changing one answer-critical query condition while holding the document and candidate evidence fixed, so evidence that supports the original query becomes invalidated for the edited query. On this diagnostic, the tested baselines retain 55.3% to 72.1% of the invalidated evidence. We introduce ConSieve, a constraint-sensitive extractive compressor trained with counterfactual supervision. On the predominantly temporal diagnostic, ConSieve retains more supporting evidence under the original query than every tested baseline and prunes 87.2% of the same evidence after the edit makes it invalid. A matched ablation links this behavior to counterfactual supervision, reducing invalid retention from 56.1% to 11.4%. Under a shared 256-token context cap in end-to-end RAG, ConSieve reaches 0.631 downstream QA F1 compared with 0.608 for the strongest tested baseline. A separate multi-budget analysis shows the same overall advantage across the tested context budgets.

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.

1 INTRODUCTION

Suppose a company archive states that Alex served as chief executive from 2013 to 2014. That sentence answers *Who was the CEO in 2013?* Changing only *2013* to *2015* barely alters the topic or wording of the query, yet the same sentence no longer supports the answer. A compressor driven mainly by semantic relevance will tend to keep evidence that looks highly relevant but violates an answer-critical condition. This simple case exposes a gap between semantic relevance and query-conditioned evidence validity. Figure 1 illustrates this distinction.

This gap matters directly for context compression. A compressor must spend a limited context budget on the evidence most useful for answering the query. Keeping text that matches the topic but violates a key condition consumes this budget, can displace valid evidence, and may steer the generator toward an incorrect answer. Existing compressors select content using signals such as semantic relevance, token importance, or estimated downstream utility (Jiang et al., 2024; Chirkova et al., 2025; Hwang et al., 2025). These signals are effective for removing broadly irrelevant text, but they do not explicitly test whether retained evidence remains valid under the query's answer-critical conditions. We study this missing capability.

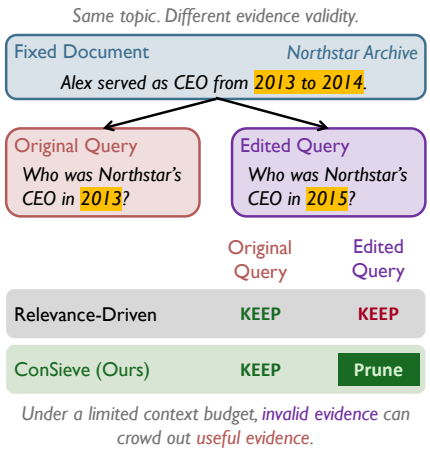


Figure 1: One answer-critical query-condition change leaves the document and previously supporting evidence fixed. Semantic relevance remains high, but constraint validity changes, so relevance-only selection can retain evidence that ConSieve should prune.

We first ask whether current compressors are sensitive to changes in evidence validity. To isolate this behavior, we construct fixed-evidence counterfactual pairs. Starting from a query supported by a document, we modify one answer-critical condition while keeping the document and candidate sentences unchanged, and retain pairs in which the edited query is no longer supported by the document. We call the supporting sentences under the original query the *old evidence*, and the same sentences after the edit the *invalidated evidence*. Because the candidate evidence is held fixed, changes in selection can be attributed directly to the query-condition change rather than to retrieval or document content. This diagnostic also exposes behavior that downstream answer accuracy can obscure, since a generator may recover from other retrieved documents or from parametric knowledge. On this diagnostic, the tested baselines keep 55.3% to 72.1% of the invalidated evidence after the query condition changes.

This failure motivates ConSieve, a sentence-level compressor whose supervision is built from the diagnosis. For each supported query and document pair, we create a counterfactual query by changing one answer-critical condition and pair it with the same document. The original query teaches the model which sentences to keep, while the edited query teaches it to prune the same sentences once they become invalid. ConSieve learns these decisions with a lightweight ModernBERT encoder (Warner et al., 2025) and a Provence-style pruning head (Chirkova et al., 2025), producing sentence scores for budgeted extraction. At inference, it operates directly on the query and retrieved text without an explicit constraint parser or external verifier.

ConSieve is designed to address this failure with counterfactual supervision. On the fixed-evidence diagnostic, it keeps more supporting evidence under the original query and prunes 87.2% of the invalidated evidence. In the matched ablation, replacing counterfactual examples with their supported counterparts raises invalid retention from 11.4% to 56.1%, linking the supervision signal to the intended selection behavior. Under the shared 256-token context budget, ConSieve reaches a downstream QA F1 of 0.631 compared with 0.608 for the strongest tested baseline. Additional non-temporal and same-document analyses characterize how this behavior transfers beyond the primary setting.

Our contributions are:

1. **A controlled diagnosis of constraint sensitivity in context compression.** We introduce a fixed-evidence counterfactual diagnostic that changes an answer-critical query condition while holding the document and candidate evidence unchanged. This design isolates whether a compressor responds to changes in evidence validity rather than to changes in retrieval or document content.
2. **A diagnostic-driven training recipe for constraint-sensitive evidence selection.** We introduce ConSieve, which uses targeted counterfactual supervision to address the failure exposed by the diagnostic. A matched ablation isolates the effect of this supervision and shows a substantial reduction in the retention of invalidated evidence.
3. **Evidence that constraint-sensitive selection matters for practical RAG.** Under shared context budgets, ConSieve improves downstream QA over the tested post-retrieval compressors. Multi-budget, non-temporal, and same-document evidence-switching analyses further characterize when the learned behavior transfers.

2 Related Work

Post-retrieval context compression. Existing post-retrieval compressors estimate which retrieved content to preserve through relevance, utility, token or sentence importance, redundancy, or downstream context-use signals. Extractive methods include Provence (Chirkova et al., 2025), LooComp (Do et al., 2026), EXIT (Hwang et al., 2025), and Sentinel (Zhang et al., 2026a). LongLLMLingua uses query-aware prompt compression and context reordering (Jiang et al., 2024); AttentionRAG uses attention-guided pruning (Fang et al., 2025); RECOMP studies compression and selective augmentation (Xu et al., 2024); PISCO learns compressed representations (Louis et al., 2025); and Ext2Gen aligns extraction with generation (Song et al., 2026). COMI uses coarse-to-fine relevance and marginal information gain to reduce redundancy (Tang et al., 2026), while OSCAR performs query-dependent online soft compression (Louis et al., 2026). These works address complementary

questions about compression efficiency, utility, and representation. We examine how a compressor's treatment of unchanged evidence responds to an answer-critical query-condition change.

Constraint and counterfactual evidence. Temporal retrieval separates temporal fit from semantic matching (Son & Oh, 2023; Zhang et al., 2025). Counterfactual RAG and evidence criticism study conflicting or perturbed evidence (Qin et al., 2026; Liu et al., 2026). DocReRank generates similar unanswerable queries for page-level reranking (Wasserman et al., 2025), and CausalNeg synthesizes retrieval negatives that violate query requirements (Zhang et al., 2026b). Contrast sets (Gardner et al., 2020) and counterfactual data augmentation (Kaushik et al., 2020) provide broader precedents. ConSieve applies same-document query negatives to post-retrieval sentence selection and evaluates old-evidence retention separately from budgeted answer quality. The fixed-evidence diagnostic guides the supervision target, and the matched ablation measures its effect on sentence selection.

3 Measuring Query-Condition Sensitivity

3.1 Relevance and constraint validity

A retriever returns ordered documents $D_q = (d_1, \ldots, d_k)$ for query q. We split each document into candidate sentences. A compressor assigns each sentence s_j a keep score $f_\theta(q, s_j; d) \in [0, 1]$, using the query and its document context. We abbreviate this score as $f_\theta(q, s_j)$ when we hold the document fixed. Selected sentences form a shorter context for the generator. We study this post-retrieval step with retrieval held constant.

A sentence provides *constraint-valid evidence* when it supports the requested answer in the evaluated document and satisfies the query's answer-critical conditions. Semantic relevance captures topical compatibility. The two can correlate, but topical compatibility alone does not establish that a sentence supports an answer under an entity, temporal, numeric, or lexical condition.

3.2 Fixed-evidence diagnostic

Consider an original query q^+, its supporting sentences, and an edit T that changes an answer-critical condition to produce $q^- = T(q^+)$. We keep the document and candidate sentences fixed and call the sentences supporting q^+ the *old evidence*. The intended diagnostic requires document support for q^+ and no document support for q^-. We call the old evidence under the edited query *invalidated evidence*.

For pair r, let $\mathcal{E}_r$ denote the set of old-evidence sentence indices. Let $k_\theta(q, s) \in \{0, 1\}$ denote a keep decision under a method's fixed threshold and selection policy (Section 5.1). Invalid retention measures how often the selector keeps invalidated evidence under q^-:

$$\text{Invalid Ret.} = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{|\mathcal{E}_r|} \sum_{j \in \mathcal{E}_r} k_\theta(q_r^-, s_j).$$

We also measure evidence selection under the supported query using q^+ F1, the sentence-level F1 against the annotated evidence, pooled over documents. The mean number of sentences kept under q^-, denoted q^- Retained Sent., measures how much text the selector retains from a document labeled unanswerable. Pair Discrimination measures whether each old-evidence sentence scores higher for the supported query:

$$\text{Pair Disc.} = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{|\mathcal{E}_r|} \sum_{j \in \mathcal{E}_r} \mathbf{1}\left\{ f_\theta(q_r^+, s_j) > f_\theta(q_r^-, s_j) \right\}.$$

Invalid retention and Pair Discrimination average over old-evidence sentences within each pair and then over pairs. Pair Discrimination counts ties as failures and does not depend on a threshold. A high value does not necessarily imply low invalid retention: a small score decrease can leave both scores above the keep threshold. Appendix B.2 defines the additional recall and ranking metrics.

This design tests rejection of invalidated evidence when the document becomes unanswerable. Appendix E separately tests evidence switching between two answerable queries with different supporting sentences.

3.3 DIAGNOSTIC CONSTRUCTION

We generate 1,646 candidate same-document pairs from StreamingQA and TimeQA evaluation rows (Liska et al., 2022; Chen et al., 2021). A blinded language-model annotator identified supporting sentences without seeing the expected edit status. The filters selected 412 pairs with cited support for q^+, no identified support for q^-, and a changed query condition. They comprise 371 StreamingQA and 41 TimeQA pairs, mostly temporal edits, with 692 old-evidence sentences.

Both members of a StreamingQA pair include the reference-date prefix needed to interpret relative time expressions. We fix pair identities and evidence indices before model evaluation. Appendix B reports the annotation procedure and a manual audit of all selected pairs.

4 CONSIEVE

ConSieve scores sentences in retrieved documents for budgeted extraction. It uses a 315M ModernBERT query–document encoder (Warner et al., 2025), a Provence-style token-pruning head (Chirkova et al., 2025), and counterfactual supervision. Figure 2 summarizes the training and inference pipeline.

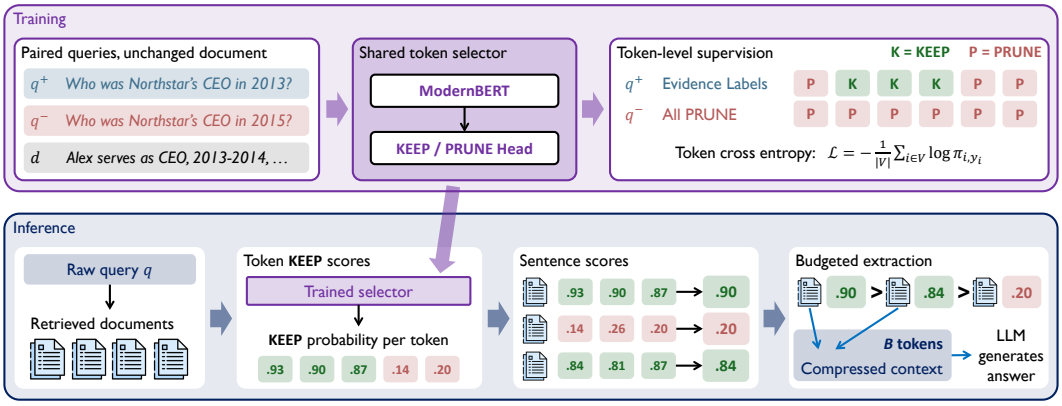


Figure 2: ConSieve training and inference. Training retains supported q^+ rows and adds generated q^- rows over the same document, assigning PRUNE targets to all supervised document tokens in the counterfactual row. Token KEEP/PRUNE scores yield sentence scores for budgeted extraction.

4.1 SENTENCE-LEVEL EVIDENCE SCORES

The encoder maps the tokenized query–document pair to contextual representations $\{\mathbf{h}_i\}_{i=1}^{L}$. A two-way pruning head produces logits $\mathbf{z}_i \in \mathbb{R}^2$ and keep probabilities:

$$\boldsymbol{\pi}_i = \mathrm{softmax}(\mathbf{z}_i), \qquad p_i = \pi_{i,\mathrm{keep}}.$$

For sentence s_j, let $\mathcal{I}_j$ contain only document-token positions whose character offsets overlap its span. Its score is

$$f_\theta(q, s_j) = \begin{cases} |\mathcal{I}_j|^{-1} \sum_{i \in \mathcal{I}_j} p_i, & \mathcal{I}_j \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases}$$

Document-segment masking excludes query and special tokens even when their offsets resemble document offsets. The selection unit is a sentence, although its score uses contextual token representations from the document. Appendix D.1 describes long-input coverage.

At a thresholded operating point, ConSieve retains s_j when $f_\theta(q, s_j) > \tau$, with τ selected using validation evidence F1 (Section 5.1). Under a hard token cap, it instead ranks all sentences without applying that threshold. A common packing rule selects whole sentences within the budget and restores document and sentence order (Appendix D.2).

4.2 COUNTERFACTUAL SUPERVISION

Starting from an evidence-supported query q^+, a language model changes one answer-critical condition while preserving the topic, producing a candidate $q^- = T(q^+)$. The edit is designed so that the unchanged document no longer supports an answer to q^-. Training retains the supported row and pairs the unchanged document with the candidate query. Every supervised document token in that counterfactual row receives a PRUNE target.

This document-level supervision provides a query-conditioned negative without requiring sentence-level counterfactual evidence annotations. Supported training rows retain their query-specific evidence labels. Appendix E separately examines whether the resulting model can redirect its selection when two answerable queries require different evidence from the same document. At inference, ConSieve receives only the query and retrieved text and does not require an explicit constraint parser, counterfactual generator, or external verifier.

4.3 TRAINING OBJECTIVE

Binary sentence annotations define evidence spans. The collator aligns these spans to document tokens and assigns $y_i \in \{0, 1\}$, where 1 denotes KEEP and 0 denotes PRUNE. Query, special, padding, and ignored positions receive label -100 and contribute no loss. For the remaining token positions $\mathcal{V}$, the pruning objective is

$$\mathcal{L}_{\text{prune}} = -\frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \log \pi_{i,y_i}.$$

Training uses only this token-pruning loss. The training set contains 22,724 StreamingQA/TimeQA rows, including generated counterfactuals. We preserve query reference dates and select the model with the lowest validation loss. Appendix A.1 gives the training settings and the matched ablation design.

5 EXPERIMENTAL DESIGN

Our evaluation separates the fixed-evidence diagnostic, the effect of counterfactual supervision, non-temporal transfer, and practical QA under shared context budgets. Bold and underlining mark the best and second-best point estimates within each comparison group.

5.1 FIXED-EVIDENCE DIAGNOSTIC

All methods score the same documents and sentences in the 412-pair diagnostic. We compare ConSieve with OpenProvence (Tateno, 2025), EXIT (Hwang et al., 2025), and Sentinel (Zhang et al., 2026a).

Sentinel uses its native inference policy; the following threshold-selection procedure applies to the other methods. Sentence-level evidence F1 on validation data (2,428 rows, 55,480 sentences) selects thresholds from $0.05, 0.10, \ldots, 0.80$, with the lower threshold breaking ties. OpenProvence and ConSieve use strict thresholds of 0.20 and 0.45. EXIT uses an inclusive threshold of 0.15. Appendix D.3 specifies baseline inference rules. Thresholds are selected exclusively on validation data.

The main tables report q^+ sentence F1, mean retained sentences under q^-, invalid retention, and strict Pair Discrimination (Section 3.2). Together, they measure useful evidence selection, residual text under an unsupported query, and response to the query change. The appendix reports q^+ evidence recall and area under the receiver operating characteristic curve (AUROC).

5.2 MATCHED ABLATION

The two ablation arms share initialization, example order and weighting, optimization, validation, and checkpoint selection. Each contains 22,590 training rows and runs for 748 optimizer steps. At 2,605 corresponding training-example positions, the counterfactual (CF) arm uses generated q^- queries with document-level PRUNE labels; the no-CF arm uses the source q^+ queries and their original labels. The remaining 134 positions lack recoverable original training labels and are excluded from

both arms. This design isolates the joint effect of changing the query and its supervision target while keeping the matched training population and schedule fixed.

Both arms select their checkpoint by validation loss and use a validation-selected keep threshold of 0.40. The main model uses the full training population; the ablation uses the matched subset so that both arms have the same source population and training schedule. Appendix A.1 provides implementation details.

5.3 HUMAN-WRITTEN NON-TEMPORAL DIAGNOSTIC

We join released UnAnsQ IDs (Zhu et al., 2019) to SQuAD 2.0 (Rajpurkar et al., 2018). An answerable query and a human-written unanswerable query share a paragraph; the original answer span matches the unanswerable query's plausible-answer span. The sentence containing that span is the old-evidence *pivot*. Human SQuAD2-CR edit labels (Lee et al., 2020) identify entity, antonym, negation, and number families. A deterministic filter excludes temporal number changes. Exact-overlap exclusions and a paragraph-disjoint split leave 1,377 calibration pairs and 3,315 test pairs over 2,366 test paragraphs. Neither SQuAD nor UnAnsQ supplies ConSieve training examples.

Each method selects the highest threshold attaining at least 95% pivot recall on calibration data and transfers it unchanged to test. We report invalid retention, realized test recall, strict Pair Discrimination, and pooled AUROC. Paired 95% percentile confidence intervals use 10,000 resamples of paragraph IDs. These intervals condition on the fitted models and calibration thresholds.

5.4 PRACTICAL RAG EVALUATION

MAIN1500, our 1,500-query QA set held out from training, contains 1,294 StreamingQA and 206 TimeQA queries. GT406, a 406-query evidence-annotated set, supports the multi-budget analysis. Appendix A reports their partial overlap and provenance.

All practical runs use one BM25 index (Robertson & Walker, 1994) with $k_1 = 0.9$, $b = 0.4$, and identical ordered top-25 documents, so the comparison isolates post-retrieval compression rather than retrieval differences. We compare OpenProvence, EXIT, LongLLMLingua (Jiang et al., 2024), Sentinel, and ConSieve. A validation-only procedure fixes candidate caps of 64, 128, 256, 512, and 1,024 standardized context tokens and selects 256 for the primary comparison.

We count standardized context tokens with the pinned `Qwen/Qwen3-4B-Instruct-2507` tokenizer (Qwen Team, 2025). The same `qwen-plus` prompt answers every compressed context, and a separate fixed prompt to the same hosted model supplies the yes/no Judge. We report token F1, exact match (EM), Judge accuracy, and mean standardized context length. GT406 also supports sentence evidence metrics. Practical intervals resample query IDs 10,000 times. Appendix D.2 records budget selection, Appendix D.3 gives failure handling and baseline adaptations, and Appendix D.5 reports efficiency measurement.

6 RESULTS

6.1 CONSIEVE REJECTS MORE INVALIDATED EVIDENCE

ConSieve improves evidence selection under the supported query while retaining less text after the query change (Table 1). Its q^+ sentence F1 is 0.744, compared with 0.668 for OpenProvence. Under the edited query, it retains an average of 0.325 sentences per document and 0.128 of the old evidence, versus 1.718 and 0.721 for OpenProvence. Pair Discrimination reaches 0.979.

Table 1: Fixed-evidence diagnostic on 412 pairs. q^+ F1 is sentence evidence F1; q^- Retained Sent. is mean retained sentences per document. Bold marks the best value and underlining the second best.

Method	q^+ F1 ↑	q^- Retained Sent. ↓	Invalid Ret. ↓	Pair Disc. ↑
OpenProvence	<u>0.668</u>	<u>1.718</u>	0.721	<u>0.907</u>
EXIT	0.574	1.879	0.717	0.864
Sentinel	0.499	1.738	<u>0.553</u>	0.766
ConSieve	**0.744**	**0.325**	**0.128**	**0.979**

The magnitude of the improvement varies across datasets, with larger gains on StreamingQA than on TimeQA. Appendix B.3 reports the dataset-specific recall, retention, Pair Discrimination, and AUROC results. Appendix E separately examines whether ConSieve can redirect its selection when two answerable queries require different supporting evidence from the same document.

6.2 Counterfactual supervision changes retention

Our matched ablation shows that counterfactual supervision directly changes how the model treats evidence after a query-condition change (Table 2). Replacing repeated supported examples with counterfactual training examples reduces invalid retention from 0.561 to 0.114 and lowers the mean number of retained sentences under q^- from 1.053 to 0.296. Pair Discrimination increases from 0.973 to 0.983, while q^+ sentence F1 changes from 0.743 to 0.728. The dominant effect of counterfactual supervision is therefore a substantially stronger rejection of evidence that becomes invalid after the query condition changes.

Table 2: Matched counterfactual-supervision ablation on the 412-pair diagnostic. Both arms use the same matched source population and training schedule.

Training arm	q^+ F1 ↑	q^- Retained Sent. ↓	Invalid Ret. ↓	Pair Disc. ↑
w/o counterfactual supervision	**0.743**	1.053	0.561	0.973
w/ counterfactual supervision	0.728	**0.296**	**0.114**	**0.983**

The same counterfactual arm also yields higher downstream point estimates on GT406 under the 256-token cap. QA F1 increases from 0.606 to 0.622, EM from 0.443 to 0.466, and Judge accuracy from 0.618 to 0.640 (Table 10). Together with the fixed-evidence results, these observations connect the training intervention to both the intended selection behavior and downstream answer quality.

6.3 Non-temporal transfer depends on edit family

We next test whether the learned behavior transfers beyond the predominantly temporal counterfactual patterns used in training. On the human-written non-temporal diagnostic, ConSieve and OpenProvence both realize q^+ recall of 0.943 after calibration. Their overall invalid-retention rates are 0.557 and 0.545, respectively, giving a ConSieve-minus-OpenProvence difference of 0.011 with a 95% confidence interval of [-0.005, 0.027].

Transfer varies substantially across edit families. ConSieve retains less invalidated evidence than OpenProvence for entity edits, with rates of 0.444 versus 0.517, and for non-temporal number edits, with rates of 0.341 versus 0.506. For antonym and negation edits, the direction reverses. These results indicate that the learned constraint sensitivity transfers beyond temporal edits for some constraint families but is not uniform across transformation types. Table 7 reports the complete family-level recall, invalid-retention, Pair Discrimination, and AUROC results.

6.4 ConSieve remains effective under matched budgets

ConSieve gives the highest F1, EM, and Judge point estimates among the five tested compressors under the shared per-query 256-token cap (Table 3). It reaches an F1 of 0.631, compared with 0.608

for OpenProvence, with a paired difference of 0.0239 and a 95% confidence interval of [0.0093, 0.0384]. The corresponding Judge difference is 0.0147 with a 95% confidence interval of [-0.0027, 0.0320].

Table 3: MAIN1500 RAG at a shared per-query 256-token cap. All methods use identical BM25@25 retrieval and generator.

Method	Answer quality			Retained context
	F1 ↑	EM ↑	Judge ↑	Mean tokens ↓
OpenProvence	0.608	0.466	0.615	224.9
EXIT	0.560	0.421	0.557	226.4
Sentinel	0.373	0.246	0.354	**222.6**
LongLLMLingua	0.422	0.267	0.384	244.1
ConSieve	**0.631**	**0.481**	**0.629**	226.2

Relative to the other tested compressors, ConSieve improves F1 by 0.0718 ([0.0521, 0.0915]) over EXIT, 0.2091 ([0.1860, 0.2330]) over LongLLMLingua, and 0.259 ([0.235, 0.283]) over Sentinel. On the evidence-annotated GT406 set, ConSieve has the highest F1 among the four plotted methods at each of the five nominal budget settings (Figure 3).

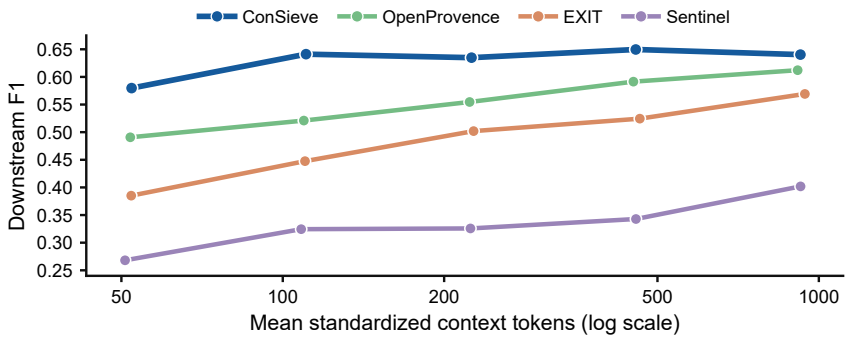


Figure 3: Budget sensitivity on GT406. Downstream F1 versus actual mean standardized context tokens under BM25@25 for OpenProvence, ConSieve, EXIT, and Sentinel, each evaluated at nominal caps of 64, 128, 256, 512, and 1,024 tokens.

ConSieve also remains lightweight relative to the decoder-based compressors. Under the sequential NVIDIA H20 protocol, ConSieve contains 315.2M parameters and has a median compressor-only latency of 3.27 seconds per query. This is comparable to OpenProvence and substantially lower than the 106.64 seconds for EXIT and the 34.07 seconds for LongLLMLingua at the 256-token operating point. Tables 8 and 9 report the natural operating points and full efficiency comparison, and Appendix D.3 describes the baseline adaptations.

7 Limitations

The main limitation of ConSieve is uneven generalization across constraint types. Counterfactual supervision yields strong gains on the patterns emphasized during training, but transfer to human-written non-temporal edits is mixed, with improvements varying substantially across edit families. This suggests that the learned constraint sensitivity remains partly dependent on the supervision distribution. Extending training to more diverse constraint transformations and domains may improve robustness. In addition, the reported inference efficiency is measured under a controlled hardware and implementation setup, and deployment-time latency may vary across systems.

8 CONCLUSION

We introduced ConSieve, a constraint-sensitive extractive compressor, together with a fixed-evidence diagnostic for measuring whether post-retrieval compressors respond to changes in evidence validity. ConSieve uses counterfactual supervision to learn when evidence that remains semantically relevant should no longer be retained after an answer-critical query condition changes. The diagnostic and matched ablation show that this supervision substantially strengthens rejection of invalidated evidence, while practical RAG experiments show corresponding gains in downstream answer quality under shared context budgets. Additional non-temporal and same-document analyses further characterize how the learned behavior transfers beyond the primary temporal setting. These results motivate treating query-conditioned evidence validity as a distinct objective in RAG context compression.

AI USE STATEMENT

We used generative AI tools to generate candidate counterfactual queries, perform the disclosed automated evidence annotation, and assist with code implementation, debugging, and experiment development. AI-based writing tools were used to improve the clarity and presentation of the manuscript. The authors verified AI-assisted code and artifacts, checked the reported results and citations, and take full responsibility for the correctness of the methods, results, and writing.

REPRODUCIBILITY STATEMENT

Sections 4–5 and the appendix describe the model, data construction, evaluation sets, ablation protocol, token-budget procedure, and baseline adaptations.

REFERENCES

Wenhu Chen, Xinyi Wang, and William Yang Wang. A dataset for answering time-sensitive questions. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://openreview.net/forum?id=9-LSfSU74n-.

Nadezhda Chirkova, Thibault Formal, Vassilina Nikoulina, and Stephane Clinchant. Provence: Efficient and robust context pruning for retrieval-augmented generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TDy5Ih78b4.

Thao Do, Dinh Phu Tran, An Vo, Seon Kwon Kim, and Daeyoung Kim. LooComp: Leverage leave-one-out strategy to encoder-only transformer for efficient query-aware context compression. *arXiv preprint arXiv:2603.09222*, 2026. URL https://arxiv.org/abs/2603.09222.

Yixiong Fang, Tianran Sun, Yuling Shi, and Xiaodong Gu. AttentionRAG: Attention-guided context pruning in retrieval-augmented generation. *arXiv preprint arXiv:2503.10720*, 2025. URL https://arxiv.org/abs/2503.10720.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4895–4924, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.253. URL https://aclanthology.org/2025.findings-acl.253/.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.91. URL https://aclanthology.org/2024.acl-long.91/.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sklgs0NFvr.

Gyeongbok Lee, Seung-won Hwang, and Hyunsouk Cho. SQuAD2-CR: Semi-supervised annotation for cause and rationales for unanswerability in SQuAD 2.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 5425–5432, 2020. URL https://aclanthology.org/2020.lrec-1.667/.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-Mcmahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13604–13622, 2022. URL https://proceedings.mlr.press/v162/liska22a.html.

Peiyang Liu, Qiang Yan, Ziqiang Cui, Di Liang, Xi Wang, and Wei Ye. Beyond semantic relevance: Counterfactual risk minimization for robust retrieval-augmented generation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2026. doi: 10.1145/3805712.3809631. URL https://doi.org/10.1145/3805712.3809631.

Edward Loper and Steven Bird. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pp. 63–70, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1118108.1118117. URL https://aclanthology.org/W02-0109/.

Maxime Louis, Hervé Déjean, and Stéphane Clinchant. PISCO: Pretty simple compression for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15506–15521, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.800. URL https://aclanthology.org/2025.findings-acl.800/.

Maxime Louis, Thibault Formal, Hervé Déjean, and Stéphane Clinchant. OSCAR: Online soft compression for RAG. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ideKAUWvFE.

Huaiyu Qin, Chunyu Wei, Yueguo Chen, and Yunhai Wang. Counterfactual reasoning for retrieval-augmented generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9U51rOnGko.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124/.

Stephen E. Robertson and Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994. doi: 10.1007/978-1-4471-2099-5_24. URL https://doi.org/10.1007/978-1-4471-2099-5_24.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4596–4604. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/shazeer18a.html.

Jungbin Son and Alice Oh. Time-aware representation learning for time-sensitive question answering. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 70–77, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.6. URL https://aclanthology.org/2023.findings-emnlp.6/.

Hwanjun Song, Jeonghwan Choi, and Minseok Kim. Aligning extraction and generation for robust retrieval-augmented generation. In *Proceedings of WSDM 2026*, pp. 596–606, 2026. doi: 10.1145/3773966.3777939.

Jiwei Tang, Shilei Liu, Zhicheng Zhang, Yujin Yuan, Libin Zheng, Wenbo Su, and Bo Zheng. Comi: Coarse-to-fine context compression via marginal information gain. In *Proceedings of ICLR 2026*, 2026. URL https://openreview.net/forum?id=OGDIXDfaN4.

Yuichi Tateno. Openprovence: An open-source implementation of efficient and robust context pruning for retrieval-augmented generation, 2025. URL https://github.com/hotchpotch/open_provence.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of ACL 2025 (Long Papers)*, pp. 2526–2547, 2025. doi: 10.18653/v1/2025.acl-long.127.

Navve Wasserman, Oliver Heinimann, Yuval Golbari, Tal Zimbalist, Eli Schwartz, and Michal Irani. DocReRank: Single-page hard negative query generation for training multi-modal RAG rerankers. In *Proceedings of EMNLP 2025 (Main Conference)*, pp. 8640–8658, 2025. doi: 10.18653/v1/2025.emnlp-main.436.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mlJLVigNHp.

Siyue Zhang, Yuxiang Xue, Yiming Zhang, Xiaobao Wu, Anh Tuan Luu, and Chen Zhao. MRAG: A modular retrieval framework for time-sensitive question answering. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3080–3118, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.167. URL https://aclanthology.org/2025.findings-emnlp.167/.

Yong Zhang, Heng Li, Yanwen Huang, Ning Cheng, Yang Guo, Yun Zhu, Yanmeng Wang, Shaojun Wang, and Jing Xiao. Sentinel: Decoding context utilization via attention probing for efficient llm context compression, 2026a. URL https://arxiv.org/abs/2505.23277.

Zhicheng Zhang, Jiwei Tang, Kuicai Dong, Xiaopeng Li, Jieming Zhu, Jingyu Li, Qianhui Zhu, Fengyuan Lu, Wang Jiaheng, Gang Wang, Hai-Tao Zheng, and Zhaocheng Du. When hard negatives hurt: Bridging the generative-discriminative gap in hard negative synthesis for retrieval. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026b. doi: 10.1145/3770855.3818118. URL https://arxiv.org/abs/2606.01304.

Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, and Ting Liu. Learning to ask unanswerable questions for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4238–4248, 2019. doi: 10.18653/v1/P19-1415. URL https://aclanthology.org/P19-1415/.

A Data and Implementation Details

A.1 Training, matched Ablation, and checkpoint selection

ConSieve and its ablation controls initialize from OpenProvence Large (Tateno, 2025) and contain 315,204,867 parameters. We select the model with the lowest validation loss. Table 4 lists the populations and selected steps.

Table 4: Training populations and checkpoint selection. Weighted rows count source-weighted examples per epoch.

Run	Raw rows	Weighted rows	Total steps	Best step
ConSieve	22,724	24,024	752	300
With counterfactual supervision	22,590	23,880	748	350
Without counterfactual supervision	22,590	23,880	748	350

Training uses two epochs, batch size 4 with gradient accumulation 16, Adafactor (Shazeer & Stern, 2018) at 5×10^{-5}, cosine decay, 10% warmup, weight decay 0.01, maximum length 4,096, and BF16. TimeQA examples have weight two and StreamingQA examples weight one. We evaluate and save every 50 optimizer steps. Each configuration uses random seed 42 and one RTX 5090. Validation evidence F1 selects inference thresholds separately from checkpoint selection.

The matched ablation uses 2,605 corresponding training-example positions for which the source q^+ queries and their original training labels are recoverable. The remaining 134 positions lack recoverable original training labels and are excluded from both arms. Source weights and example order are identical across arms. The main model uses the full training population and a validation-selected threshold of 0.45; both ablation arms use the matched subset and a threshold of 0.40.

A.2 Data, overlap, and retrieval coverage

StreamingQA queries retain the `Today is` reference-date prefix on both sides of a pair. The data keep their train, validation, and evaluation assignments. GT406 contains 406 queries, including 200 StreamingQA and 206 TimeQA examples. The diagnostic's original-query IDs overlap GT406 in 53 cases and MAIN1500 in 119 cases. No counterfactual-query ID overlaps either practical set. MAIN1500 contains 1,294 StreamingQA and 206 TimeQA queries; 233 query IDs overlap GT406, comprising 27 StreamingQA and all 206 TimeQA queries. These sets share evaluation examples. Overlap checks use exact IDs and normalized text; they do not attempt to remove semantic or near-duplicate matches.

The retrieval corpus uses the WMT News Crawl 2007–2020 documents identified by StreamingQA's official sorting-key whitelist and one document per unique TimeQA evaluation context. After duplicate-ID resolution, it contains 11,393,471 WMT documents and 1,434 TimeQA documents, indexed with Lucene. GT406 uses 10,150 ordered BM25@25 document occurrences and 427,536 sentence occurrences. Retrieval finds the gold document for 299 queries and misses it for 107; 205 of the 958 original positive evidence units are absent from retrieval. Gold-evidence recall retains these misses in its denominator. Unannotated distractor sentences remain unlabeled.

A.3 Generation and cached outputs

The hosted `qwen-plus` generator uses temperature 0, top-p 1, and a 128-token output limit. A separate temperature-zero Judge prompt uses an 8-token output limit. It checks reference answers separately and accepts a completed `Yes` judgment for any reference. We cache exact requests and responses and permit at most three explicit attempts. All reported GT406 panels contain 406 completed answers and judgments. Because `qwen-plus` is a mutable alias, the cached outputs and their evaluation dates identify each evaluated deployment. We ran the MAIN1500 practical comparison on August 13–15, 2026 (UTC).

B Fixed-Evidence Diagnostic Details

B.1 Annotation and manual audit

For the construction in Section 3.3, the annotator receives stable sentence indices and randomized question order without access to the expected support status of either query. Pair identities and evidence indices are fixed before model evaluation.

We additionally manually audited all 412 selected diagnostic pairs to assess whether the edit produces the intended support transition. The audit confirmed that the original query is supported by the document while the counterfactual query is unsupported for 395 of the 412 pairs, corresponding to 95.9% of the diagnostic. We keep the diagnostic population fixed across all reported methods so that every compressor is evaluated on exactly the same pairs and evidence indices.

B.2 Additional metrics

Positive-evidence recall (q^+ recall) replaces q^- with q^+ in the invalid-retention expression in Section 3.2. For q^+ sentence F1, we pool true positives, false positives, and false negatives across documents and compute $2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$. Table 5 reports the additional metrics for the main comparison and the matched ablation separately.

Table 5: Additional fixed-evidence diagnostic metrics for the main comparison and matched ablation.

Method	q^+ **Recall** ↑	**AUROC** ↑
Main comparison		
OpenProvence	<u>0.854</u>	<u>0.626</u>
EXIT	0.814	0.570
Sentinel	0.653	0.534
ConSieve	**0.899**	**0.942**
Matched ablation		
w/o counterfactual supervision	**0.905**	0.795
w/ counterfactual supervision	0.893	**0.945**

Let $\mathcal{A}$ and $\mathcal{B}$ be the pooled original and counterfactual evidence scores. AUROC is

$$\frac{1}{|\mathcal{A}||\mathcal{B}|} \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{B}} \left[\mathbf{1}\{a > b\} + \frac{1}{2} \mathbf{1}\{a = b\} \right].$$

It compares scores across pairs and awards half credit to ties. Strict Pair Discrimination compares matching evidence within each pair and counts ties as failures.

B.3 Dataset-specific results

Table 6 separates StreamingQA and TimeQA. Lower invalid retention can accompany lower valid-evidence recall, as Sentinel's TimeQA result illustrates. These selector metrics use different populations and readouts from the downstream QA evaluation.

C Non-Temporal Diagnostic Details

The test contains 3,315 pairs, with 1,377 separate calibration pairs. Only the sentence containing the matched answer span has positive and negative labels under the two queries; other sentences remain unlabeled. We therefore evaluate that pivot without inferring exhaustive sentence F1.

Calibration selects strict thresholds of approximately 0.686562 for ConSieve and 0.494548 for OpenProvence. Table 7 reports every family. Negative ConSieve-minus-OpenProvence invalid-retention differences favor ConSieve. Confidence intervals use shared paragraph-cluster bootstrap draws, random seed 20260810, and no multiple-comparison adjustment.

Table 6: Dataset-specific diagnostic results: 371 StreamingQA pairs and 41 TimeQA pairs.

Method	q^+ **Recall** ↑	**Invalid Ret.** ↓	**Pair Disc.** ↑	**AUROC** ↑
StreamingQA (371 pairs)				
OpenProvence	0.869	0.734	0.910	0.628
EXIT	0.852	0.764	0.875	0.567
Sentinel	0.688	0.599	0.785	0.534
ConSieve	**0.915**	**0.093**	**0.997**	**0.965**
TimeQA (41 pairs)				
OpenProvence	0.713	0.602	**0.877**	0.621
EXIT	0.476	0.288	0.764	0.626
Sentinel	0.336	**0.129**	0.591	0.550
ConSieve	**0.752**	0.445	0.822	**0.700**

Table 7: SQuAD non-temporal results at calibration-selected 95% recall operating points. ΔIR is ConSieve minus OpenProvence.

A: Realized test q^+ recall

Family	ConSieve	OpenProvence
All	0.943	0.943
Entity	0.946	0.948
Antonym	0.947	0.945
Number	0.933	0.921
Negation	0.931	0.936

B: Invalid retention and paired 95% intervals

Family	N	ConSieve	OpenProvence	ΔIR [95% CI]
All	3315	0.557	**0.545**	0.011 [-0.005, 0.027]
Entity	1454	**0.444**	0.517	-0.074 [-0.096, -0.051]
Antonym	1028	0.623	**0.571**	0.052 [0.023, 0.080]
Number	164	**0.341**	0.506	-0.165 [-0.232, -0.101]
Negation	669	0.753	**0.577**	0.176 [0.142, 0.211]

C: Threshold-free ranking metrics

Family	ConSieve AUROC	OpenProvence AUROC	ConSieve Pair Disc.	OpenProvence Pair Disc.
All	0.860	**0.865**	0.934	**0.948**
Entity	**0.886**	0.871	0.936	**0.937**
Antonym	0.849	**0.859**	0.932	**0.951**
Number	**0.922**	0.868	0.970	**0.976**
Negation	0.796	**0.859**	0.924	**0.961**

The matched CF and no-CF controls realize test q^+ recall of 0.954 and 0.942 after calibration. Their invalid retention is 0.625 and 0.635. The no-CF-minus-CF invalid-retention difference is 0.010 (95% CI [-0.001, 0.021]).

D Additional Practical Results and Baseline Protocols

D.1 Sentence alignment and long documents

We segment documents with the Natural Language Toolkit (NLTK) (Loper & Bird, 2002) and apply the token-to-sentence alignment in Section 4.1. We partition documents that exceed the encoder window at sentence boundaries, assign every source sentence to one fragment, and restore selected sentences to source order. The encoder can still truncate an individual sentence that alone exceeds the available window.

D.2 Budget selection

Budget selection uses a source-balanced 256-query validation subset. Among the candidate caps in Section 5.4, we choose the cap nearest the median natural context length across the five methods, yielding 256 tokens. Under hard caps, sentence-based methods rank all candidate sentences without applying their natural thresholds and pack whole sentences in score order under a shared token-accounting rule. Selected sentences are restored to source order, and the assembled context is rechecked against the cap; sentences are never clipped. Counts include only the retrieved context, excluding the question, instructions, chat template, and answer.

D.3 Baseline adaptations

OpenProvence uses `hotchpotch/open-provence-reranker-large-v1`. We use its context-pruning output and apply the shared sentence alignment, long-document fragmentation, validation thresholding, and token-budget protocol used throughout our evaluation. EXIT uses official `EXIT` (`doubleyyh/exit-gemma-2b`); we disable its retriever/reader and chunk deterministically beyond the native window. LongLLMLingua uses official `LLMLingua` 0.2.2 (`NousResearch/Llama-2-7b-hf`); deterministic fragmentation and budget reruns preserve source coverage, with three exhausted 256-token cases assigned empty contexts. Sentinel uses official `yzhangchuck/Sentinel` with `ReRaWo/Sentinel` and a Qwen2.5-0.5B proxy.

We use Sentinel's native inference policy as the primary baseline for faithful reproduction. On the fixed-evidence diagnostic, its inclusive threshold is 0.75; if no sentence passes, it retains the first maximum-scoring sentence. EXIT's diagnostic threshold is validation-selected at 0.15; its practical compression follows the shared-budget protocol.

Baseline availability and interfaces. The comparison uses public post-retrieval compressors that accept the shared retrieved documents and produce textual context for the fixed generator under a common token budget. As of August 11, 2026, LooComp's paper-linked resources provided no working official implementation or checkpoint (Do et al., 2026). MRAG changes the retrieval framework (Zhang et al., 2025); COMI and OSCAR produce merged or soft model-side representations (Tang et al., 2026; Louis et al., 2026).

D.4 NATURAL OPERATING POINTS

Table 8: MAIN1500 natural operating points under identical retrieval, generation, Judge, and token accounting.

Method	F1 ↑	EM ↑	Judge ↑	Mean tokens ↓	Empty (%) ↓
Raw BM25@25	0.614	0.457	0.628	31,233.2	0.0
OpenProvence	0.620	0.469	0.632	848.1	3.3
EXIT	0.511	0.375	0.522	387.2	16.3
LongLLMLingua	0.522	0.346	0.491	469.6	0.0
ConSieve	0.637	0.481	0.645	1,322.4	2.7

D.5 EFFICIENCY

Table 9: Sequential single-GPU NVIDIA H20 compressor-only efficiency, omitting retrieval and generation.

Method	Params (M) ↓	Median s/q ↓	P95 s/q ↓	Peak GiB ↓
OpenProvence	315.2	3.25	6.91	1.8
Sentinel	494.0	3.99	5.55	20.9
EXIT	2,584.6	106.64	384.87	14.1
LongLL. natural	6,738.4	11.23	20.00	16.0
LongLL. @256	6,738.4	34.07	132.03	16.0
ConSieve	315.2	3.27	6.97	3.2

D.6 DOWNSTREAM ABLATION

Table 10: Matched ablation on GT406 at 256 tokens. Values are point estimates.

Training arm	**QA F1**	**EM**	**Judge**	**Mean tokens**
w/o counterfactual supervision	0.606	0.443	0.618	**227.2**
w/ counterfactual supervision	**0.622**	**0.466**	**0.640**	229.0

E QUERY-CONDITIONED EVIDENCE REDIRECTION

ConSieve redirects evidence selection between answerable questions about the same document. We check this behavior on 255 StreamingQA question pairs with disjoint annotated evidence sets, spanning 196 documents and 427 queries. We group questions by exact document identity and sentence boundaries; these pairs form the disjoint subset of 401 pairs with different evidence sets. The existing annotations provide proxy evidence labels: they need not exhaust support or establish that the other question's evidence is invalid. The questions can differ beyond a single condition.

For questions q_A, q_B, let E_A, E_B denote their annotated evidence sets, K_A, K_B the selected sentences, and f_A, f_B the sentence scores. *Rank reversal* measures whether each cross-set sentence pair reverses its ordering in the direction required by the query:

$$\text{Reversal}_{AB} = \frac{1}{|E_A||E_B|} \sum_{a \in E_A} \sum_{b \in E_B} \mathbf{1}\{f_A(a) > f_A(b) \ \wedge \ f_B(b) > f_B(a)\}\,.$$

Ties count as failures. *Correct transition* measures the fraction of evidence sentences that the selector retains for their own question and drops for the other:

$$\text{Transition}_{AB} = \frac{1}{2}\left(\frac{|(E_A \cap K_A) \setminus K_B|}{|E_A|} + \frac{|(E_B \cap K_B) \setminus K_A|}{|E_B|}\right).$$

Complete switch requires both directions to retain all own-query evidence and exclude all evidence from the other annotated set; it permits selected sentences outside both sets. Recall averages the retained fraction of own-query evidence across directions. We average all metrics equally over pairs and keep the thresholds and baseline policies from Section 5.1 unchanged.

Table 11: Query-conditioned evidence redirection on 255 same-document question pairs. All metrics refer to the annotated evidence sets.

Method	**Rank reversal ↑**	**Correct transition ↑**	**Complete switch ↑**	**Recall ↑**
OpenProvence	<u>0.923</u>	<u>0.766</u>	<u>0.349</u>	<u>0.798</u>
EXIT	0.863	0.623	0.212	0.694
Sentinel	0.838	0.650	0.235	0.680
ConSieve	**0.945**	**0.787**	**0.412**	**0.825**

ConSieve has higher point estimates than OpenProvence for rank reversal, correct transition, and complete switch (Table 11). Their paired differences are +0.0214 (95% CI [0.0006, 0.0427]), +0.0218 ([-0.0033, 0.0466]), and +0.0627 ([0.0076, 0.1179]), respectively. Intervals use 10,000 shared document-cluster bootstrap draws with random seed 20260925, without multiple-comparison adjustment. The rank reversals establish query-conditioned changes in relative evidence scores, which a uniform document-wide additive shift alone cannot produce. Together with the selection metrics, this sanity check supports evidence redirection between answerable questions in the evaluated documents.