

Paraphrase-Induced Retrieval Instability: Systematic Measurement and the Limits of Inference-Time Mitigation

Aravind Mandiga
University of Georgia
aravind.mandiga@uga.edu

Guoming Li
University of Georgia
gmli@uga.edu

Jin Lu
University of Georgia
jin.lu@uga.edu

Abstract

Dense retrieval can return different evidence when a query is paraphrased, even when the meaning stays the same. This means that linguistic variation can create uncertainty in the evidence given to a retrieval-augmented generation (RAG) system. We study this problem across HotpotQA, FEVER, and SQuAD 2.0 using five dense encoders spanning 22M to 8B parameters and multiple model families. Paraphrase Jaccard distance remains 32 to 50 percent, showing that retrieval instability persists with stronger and alternative encoders. We then ask how much this problem can be reduced at inference time. Multi-source consensus over paraphrased queries and different retrievers improves retrieval coverage. Fixed source-type weights can help in harder settings. However, adapting those weights to each query using retrieval stability changes Coverage@20 by at most ± 0.3 percentage points per setting. Retrieval stability is more useful as a label-free routing signal: on HotpotQA/MiniLM, routing only unstable queries through the stability-conditioned consensus configuration reduces retrieval calls by 47 percent for a 0.6 percentage-point Coverage@20 cost compared with always using that configuration. Overall, paraphrase-induced retrieval instability is a persistent source of uncertainty in retrieved evidence, and the query-adaptive stability component provides only a small marginal refinement.

1 Introduction

A retrieval-augmented generation (RAG) system (Lewis et al., 2020; Karpukhin et al., 2020) should ideally retrieve similar evidence for queries that ask the same thing. In practice, small wording changes can change the retrieved top- k set. As a result, two meaning-preserving versions of the same query can give the generator different evidence. We view this as a source of uncertainty caused by linguistic variation. Standard retrieval

metrics measure whether relevant evidence is retrieved, but they do not measure whether that evidence stays consistent when the query is paraphrased. We study this problem as *paraphrase-induced retrieval instability*.

We measure this instability on HotpotQA, FEVER, and SQuAD 2.0 across five dense encoders: MiniLM, Qwen3-Embedding-0.6B, Qwen3-Embedding-8B, E5-large-v2, and BGE-large-en-v1.5. Paraphrasing changes 32–50% of the top-20 retrieval set by Jaccard distance, and the effect remains across model scale and encoder family. We use $st(q)$, the average top- k overlap between the original query and its paraphrases, as a simple per-query measure of retrieval stability. It is a diagnostic of how sensitive retrieval is to wording, not a new general uncertainty metric.

We then ask what can be improved at inference time without retraining the retriever or generator. Query reformulation, dense and sparse retrieval, and rank aggregation are established methods. We use these components in a controlled mitigation study to separate source diversity, fixed source-type weighting, and query-adaptive stability weighting. Combining retrieval views improves coverage, and fixed source-type weighting can contribute substantially in harder settings. Making those weights depend on $st(q)$ changes Coverage@20 only slightly and can hurt when retrieval is already stable. In contrast, $st(q)$ is useful for deciding when the more expensive consensus step is worth running.

We make three contributions:

1. **Measurement.** We systematically measure paraphrase-induced retrieval instability across three tasks and five dense encoders, and show that it remains with stronger models and different encoder families.
2. **Robustness and characterization.** We show that the effect remains under another paraphraser and high-fidelity filtering, and that different re-

trieval views provide complementary evidence.

3. Mitigation limits and routing. Multi-source consensus improves retrieval, while the query-adaptive $\text{st}(q)$ component changes Coverage@20 by at most ± 0.3 pp per setting. Retrieval stability is more useful as a label-free routing signal; on HotpotQA/MiniLM, selective use of the full configuration reduces retrieval calls by 47% for a 0.6 pp Coverage@20 cost.

2 Related Work

Query reformulation and retrieval fusion.

Several methods improve retrieval by changing or expanding the query. HyDE generates a hypothetical document before dense retrieval (Gao et al., 2023), Query2Doc expands a query with LLM-generated pseudo-documents (Wang et al., 2023), and RAG-Fusion retrieves with several generated queries and combines the ranked lists with reciprocal rank fusion (Rackauckas, 2024). Other work learns query rewrites from downstream or ranking feedback (Ma et al., 2023; Mao et al., 2024). Dense and sparse retrieval are also commonly combined through rank fusion (Karpukhin et al., 2020; Robertson and Zaragoza, 2009; Cormack et al., 2009). These methods mainly aim to improve retrieval quality. We use paraphrases for a different purpose: to test whether equivalent wording changes the retrieved evidence, and then to test whether combining several retrieval views can reduce this problem.

Retrieval evaluation and consistency. BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2023) evaluate retrieval and embedding models across many datasets, tasks, and domains. Query performance prediction methods estimate how difficult a query is without relevance labels (Carmel and Yom-Tov, 2010). These settings measure performance or difficulty, but not whether the same information need returns similar evidence after paraphrasing. Con-RAG (Hamman et al., 2025) is closer to our setting: it measures retriever, generator, and end-to-end consistency across paraphrased queries, then trains the generator to improve consistency. RobustRAG (Xiang et al., 2024) studies a different problem, where retrieved text is changed by an adversary. Our work focuses on the retrieval stage under meaning-preserving wording changes and tests mitigation without retraining.

Uncertainty, variation, and disagreement.

Recent uncertainty-aware NLP work studies several forms of variation and disagreement. Sharma et al. (2025) study calibration of neural rankers under noisy labels, Troshin et al. (2025) compare ways to generate diverse plausible LLM responses, and Nachshoni et al. (2025) evaluate QA when evidence supports multiple or conflicting answers. Our setting is different: the information need is kept fixed, and uncertainty comes from changes in which evidence is retrieved when the query is reworded.

3 Measuring Instability and Testing Mitigation

3.1 Retrieval Stability

Let q be a query and let $R(q, K)$ be its top- K retrieved set. For m meaning-preserving paraphrases $q_1, \dots, q_m$, we define retrieval stability as the average Jaccard overlap with the original query:

$$\text{st}(q) = \frac{1}{m} \sum_{i=1}^m J(R(q, K), R(q_i, K)). \quad (1)$$

We use $m = 5$ and $K = 20$ in the main experiments. A score near 1 means that paraphrases return similar evidence, while a score near 0 means that they return very different evidence. We report $1 - \text{st}(q)$ as Jaccard distance, our main instability measure. We also report top- K replacement rate and Rank-Biased Overlap as checks.

The same overlap statistic has a different meaning when we compare different retrievers. Low overlap between two paraphrases of the same query is a reliability problem. Low overlap between dense retrieval and BM25 can be useful because the two retrievers may find different relevant evidence. We therefore use paraphrase overlap to measure instability and retriever disagreement to measure source diversity.

3.2 Consensus as an Inference-Time Mitigation

To test inference-time mitigation, we combine several query and retrieval views. For each query, we use the original form and five paraphrases. Each of the six query variants is sent to three channels: paragraph-level dense retrieval, paragraph-level BM25, and sentence-level dense retrieval mapped back to parent paragraphs. This gives 18

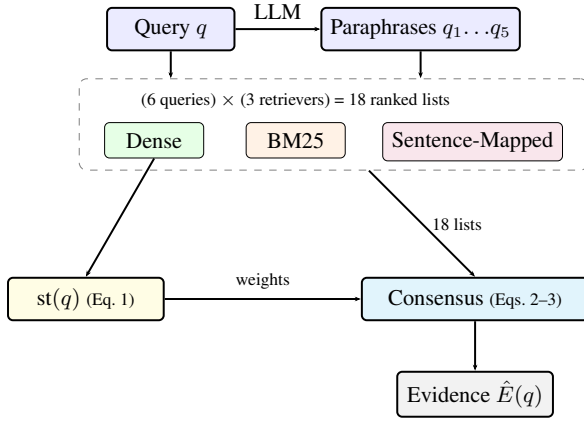


Figure 1: Consensus mitigation setup used in our decomposition: 6 query variants $\times$ 3 retrievers = 18 ranked lists. We test both fixed source weights and the stability-conditioned weighting rule in §3.3; output $\hat{E}(q)$ is the top- k set.

ranked lists, each with $K = 20$ results. We use this 18-source consensus configuration as an experimental tool for the decomposition below, not as a new retrieval primitive.

For a source s with ranked list $L_s(q)$, evidence item e receives a rank-discounted score:

$$\text{score}_s(e | q) = \sum_{r=1}^K \frac{\mathbf{1}[e = L_s(q)[r]]}{r}. \quad (2)$$

The final consensus score sums evidence across sources:

$$S(e | q) = \sum_{s \in \mathcal{S}} w_s(q) \text{score}_s(e | q). \quad (3)$$

The $1/r$ discount gives more weight to evidence that appears near the top of a source list. We compare other rank discounts in Appendix B. Consensus itself does not require $\text{st}(q)$. In the experiments, we compare equal or fixed source weights with the stability-conditioned variant described next. For shorthand in the appendix, we refer to the fully weighted, stability-conditioned configuration as Source-Consensus Evidence Retrieval (SCER). Figure 1 summarizes this setup.

3.3 Does Stability-Conditioned Weighting Help?

We next test a secondary hypothesis: whether the stability score should change the source weights for each query. The tested rule increases the dense weight as retrieval becomes more stable, decreases the BM25 weight, keeps the sentence-mapped weight fixed, and applies one discount to

all paraphrase sources. We compute $\text{st}(q)$ from dense top- K overlap. The full equations, selected weight schedule, and aggregation algorithm are given in Appendix A.

This weighting rule is one tested variant of the consensus setup. Our ablations later separate the effects of combining sources, fixed source-type weighting, and the query-adaptive $\text{st}(q)$ component; the last changes Coverage@20 by at most ± 0.3 pp per setting.

4 Experimental Setup

4.1 Datasets

We use validation questions from three Wikipedia-based benchmarks:

- HotpotQA (Yang et al., 2018): 7,405 multi-hop questions with two gold evidence paragraphs. The retrieval corpus contains 507,494 paragraphs pooled from the train and validation contexts.
- FEVER (Thorne et al., 2018): 6,506 SUPPORTS/REFUTES claims with usable evidence after preprocessing. The corpus contains 13,558 Wikipedia pages. We evaluate page-level retrieval and require Coverage@ k to contain the strict union of all annotated evidence pages.
- SQuAD 2.0 (Rajpurkar et al., 2018): 5,928 answerable questions and a corpus of 20,233 paragraphs. Each question has one gold paragraph, so Coverage@20 is equivalent to hit@20.

4.2 Models, Retrieval, and Paraphrases

The full mitigation experiments use three dense encoders: all-MiniLM-L6-v2 (22M) (Reimers and Gurevych, 2019), Qwen3-Embedding-0.6B, and Qwen3-Embedding-8B (Zhang et al., 2025). To test whether the instability result depends on these models, we also measure it with E5-large-v2 (Wang et al., 2022) and BGE-large-en-v1.5 (Xiao et al., 2023). The E5 and BGE runs are robustness checks for the measurement result; we do not run the full stability-conditioned consensus or downstream experiments with them.

For each query, Qwen3-32B-FP8 (Yang et al., 2025) generates five paraphrases in one call. We use these paraphrases as a controlled probe of wording variation, not as a complete model of natural user reformulation. NLI validation with DeBERTa-v3 (He et al., 2023) gives bidirectional entailment rates of 86.7–91.6% across the three datasets, and a GPT-5.4 (OpenAI, 2026) audit

rates 89.2–90.4% of sampled paraphrases as semantically equivalent or better. The high-fidelity-only and Llama-3.1-8B (Grattafiori et al., 2024) checks are reported in §5.2; prompts and full audit details are in Appendix A.

Dense retrieval uses exact inner-product search. Sparse retrieval uses BM25 with Pyserini (Lin et al., 2021). The sentence-mapped channel retrieves sentences with the dense encoder and maps them back to their parent paragraphs. Main mitigation experiments use $K = 20$ results per source.

4.3 Comparisons and Evaluation

We organize comparisons by what they test. Single-source baselines are Dense and BM25. Two-source fusion is Hybrid RRF (Cormack et al., 2009). Query and multi-source aggregation includes RAG-Fusion, all-source RRF, uniform consensus, and stability-conditioned consensus. We also compare with stronger retrieval models such as SPLADE and ColBERTv2 in Table 6; reranking results are in the appendix.

The main results use a fixed 20/80 calibration/test split for each benchmark (seed 42). Calibration data are used only to select source weights and the routing threshold. We use Coverage@20 as the main metric: a query succeeds only if all gold evidence units appear in the top 20. Recall, MRR, nDCG, and paired-bootstrap significance tests are reported as secondary checks. We also repeat the calibration/test split with four additional seeds; this robustness check is discussed with the mitigation results.

Implementation details, prompts, extended metrics, and reproducibility information are provided in the appendix. Code and data are available at <https://github.com/aravind-mandiga/scer>.

5 Results

5.1 How Large Is Paraphrase-Induced Retrieval Instability?

Across the three main encoders and three datasets, paraphrasing changes the dense top-20 set by 32.1–49.8% Jaccard distance. The mean fraction of top-20 documents that are actually replaced is 20.5–35.5%, and RBO ranges from 0.59 to 0.70. These measures give the same overall pattern, so the effect is not limited to one overlap metric. Table 1 reports the mean Jaccard overlap $st(q)$, while the full metric table is in Appendix B. Figure 2

Benchmark	MiniLM	Qwen3-0.6B	Qwen3-8B
HotpotQA	0.50 ± 0.16	0.50 ± 0.15	0.58 ± 0.14
FEVER	0.52 ± 0.14	0.54 ± 0.13	0.61 ± 0.13
SQuAD 2.0	0.59 ± 0.16	0.63 ± 0.16	0.68 ± 0.14

Table 1: Mean retrieval stability $st(q)$ (Jaccard@20, mean $\pm$ SD). Lower overlap means greater paraphrase-induced instability. Jaccard distance is $1 - st(q)$.

Embedder	HotpotQA	FEVER	SQuAD 2.0
E5-large-v2	39.8%	49.8%	40.3%
BGE-large-en-v1.5	40.4%	41.2%	40.3%

Table 2: Paraphrase Jaccard distance for E5-large-v2 and BGE-large-en-v1.5, used as encoder-family robustness checks. Higher values mean more instability.

shows the corresponding stability distributions.

Example. For the HotpotQA question “*What is the name of the Isles who were led by Aonghus Mór. . . ?*”, MiniLM retrieves both gold paragraphs for the original wording and for two paraphrases that keep *Isles*. It retrieves neither gold paragraph for three paraphrases that replace *Isles* with *islands*. This specific proper-noun change is rare in our fidelity audit, but it shows how a small lexical change can move retrieval to a different corpus neighborhood.

5.2 Does the Instability Persist Across Models and Paraphrasers?

Increasing model size reduces instability, but does not remove it. With Qwen3-Embedding-8B, Jaccard distance is still 32.1–41.9% across the three datasets. We also tested two additional encoders that were not used in the consensus experiments. Table 2 shows that E5-large-v2 and BGE-large-en-v1.5 also have large paraphrase-induced changes.

The result also remains when we change the paraphrase generator. With MiniLM retrieval, Llama-3.1-8B paraphrases give Jaccard distances of 54.6%, 53.0%, and 47.0% on HotpotQA, FEVER, and SQuAD 2.0, compared with 49.6%, 47.8%, and 41.0% for the Qwen3-32B paraphrases. The exact values change, but the instability remains.

Paraphrase quality does not explain the main result. When we restrict the audited sample to high-fidelity paraphrases, Jaccard distance remains 31–49% across the three datasets and main encoders. We therefore treat LLM paraphrases as a controlled way to test sensitivity to wording, while

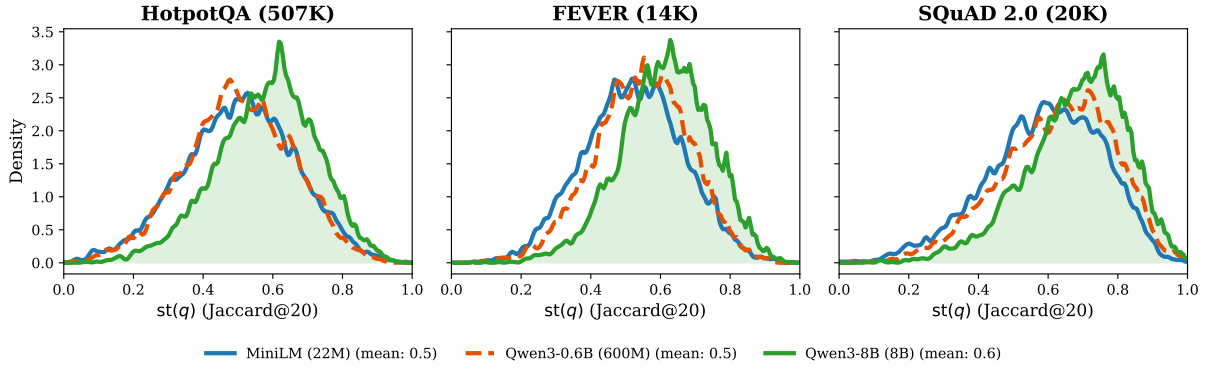


Figure 2: Distribution of $st(q)$ for the three main encoders. A score of 0 means disjoint top-20 sets and 1 means identical sets. Instability decreases with stronger encoders but remains substantial.

Dataset	Original– paraphrase word overlap	Paraphrase– paraphrase word overlap	Within-query rank correlation
HotpotQA	0.57	0.49	0.17–0.26
FEVER	0.43	0.41	0.14–0.26
SQuAD 2.0	0.39	0.30	0.20–0.28

Table 3: Content-word overlap between the original query and its paraphrases and between pairs of paraphrases. The final column gives the range across the three main encoders of within-query rank correlation between content-word distance and top-20 retrieval change.

naturally occurring user reformulations remain future work.

What kinds of wording changes matter? We further characterize which paraphrase changes are associated with retrieval instability. As shown in Table 3, greater content-word change is associated with greater top-20 change, but the relationship is modest: within-query rank correlations range from 0.14 to 0.28 across datasets and encoders. Importantly, lexical change does not explain the effect completely. On HotpotQA and SQuAD 2.0, paraphrases that preserve exactly the same set of content words, changing only their order or function words, still change 24–45% of the retrieved top-20 set. The most reworded third also shows higher gold-evidence miss rates than the least reworded third on HotpotQA (by 5.4–6.4 pp) and SQuAD 2.0 (2.0–3.4 pp); FEVER gaps are at most 0.7 pp.

Paraphrases also differ from one another at least as much as they differ from the original query in content-word overlap (Table 3). In the fidelity audit, even paraphrases rated fully faithful change 36–45% of the top-20 set, while subject-changing

paraphrases show 11–14 percentage points more retrieval change than fully faithful paraphrases. On HotpotQA, changing a name or number is associated with 2.3–3.0 additional percentage points of top-20 change and 3.7–4.4 percentage points of additional gold-evidence miss after controlling for lexical distance and query identity. These results suggest that the amount and type of linguistic variation are associated with instability, but substantial instability remains even under faithful and lexically conservative reformulations. Detailed analyses are in Appendix A.4.

5.3 What Does Multi-Source Consensus Improve?

Table 4 compares single-source retrieval, Hybrid, and the stability-conditioned 18-source consensus configuration. This configuration improves over Hybrid in all nine benchmark-encoder settings. The gain is largest on HotpotQA and with weaker encoders, ranging from +2.7 pp on HotpotQA/MiniLM to +0.2 pp in several near-ceiling settings. All nine gains are significant at uncorrected $p < 0.05$, and six survive Bonferroni correction; exact tests are in Appendix B. Cov@5, MRR, and nDCG@20 show the same ordering on HotpotQA/MiniLM.

The main question is where this gain comes from. Table 5 summarizes the decomposition, while Table 18 in Appendix B reports all nine settings. Uniform consensus already improves over all-source RRF, and removing the query-adaptive part changes Coverage@20 by at most ± 0.3 pp per setting.

Comparison with stronger retrievers. Table 6 prevents a stronger claim than the results support. On HotpotQA/MiniLM, SPLADE alone reaches

	Dense BM25 Hybrid			Stability-cond. consensus	Δ
<i>all-MiniLM-L6-v2 (22M)</i>					
HotpotQA	.372	.498	.561	.588	+2.7
FEVER	.887	.933	.945	.951	+0.6
SQuAD 2.0	.896	.944	.965	.975	+1.0
<i>Qwen3-Embedding-0.6B (600M)</i>					
HotpotQA	.507	.498	.617	.637	+2.0
FEVER	.937	.933	.952	.954	+0.2
SQuAD 2.0	.929	.944	.974	.980	+0.5
<i>Qwen3-Embedding-8B (8B)</i>					
HotpotQA	.701	.498	.721	.729	+0.8
FEVER	.956	.933	.962	.964	+0.2
SQuAD 2.0	.961	.944	.983	.985	+0.2

Table 4: Coverage@20 for the three main encoders. Δ is stability-conditioned consensus minus Hybrid in percentage points.

Configuration	HotpotQA			Mean 9 settings
	MiniLM	0.6B	8B	
All-source RRF	.553	.618	.728	.855
Uniform consensus	.570	.628	.732	.859
No $st(q)$ adaptation	.588	.634	.727	.862
Stability-cond. consensus	.588	.637	.729	.863

Table 5: Consensus decomposition. The mean is Coverage@20 across all nine benchmark-encoder settings. The complete per-setting ablation is in Table 18 of Appendix B.

.610, above BM25-based stability-conditioned consensus at .588. Replacing BM25 with SPLADE inside the same stability-conditioned consensus setup raises Coverage@20 to .617. ColBERTv2 also exceeds the MiniLM-based stability-conditioned consensus on SQuAD 2.0 (.981 vs. .975). Thus, multi-source consensus can help, but it does not always beat a stronger single retriever.

5.4 What Is Retrieval Stability Useful For?

The small query-adaptive effect does not mean that $st(q)$ is useless. Low-stability queries tend to benefit more from the full configuration, but the per-query correlation between $st(q)$ and the gain is weak (Appendix B, Figure 3 and Table 15). This makes stability more useful for coarse routing than for predicting the exact gain of each query. The decomposition shows an important distinction. Fixed source-type weighting can contribute substantially on HotpotQA, but removing only the $st(q)$ adaptation changes Coverage@20 by at most ± 0.3 pp per setting. A targeted direction check across five calibration/test splits finds that the full

tested weighting schedule improves over uniform consensus by $+0.33 \pm 0.02$ pp, while the reversed schedule improves by only $+0.02 \pm 0.02$ pp. This confirms that the selected weighting direction is better than the reverse, while Table 5 separately isolates the much smaller query-adaptive component.

The main mitigation result is also stable across data splits. Across seeds 42, 0, 1, 2, and 3, the mean stability-conditioned consensus gain over Hybrid ranges from 0.89 to 0.98 pp across the nine settings, and all nine settings remain positive on every split. Re-tuning the weights independently on each split gives a mean gain of 0.91 pp. Full split and weighting-direction results are in Appendix B.

Routing. Because $st(q)$ does not require relevance labels, it can decide when to use expensive consensus. On HotpotQA/MiniLM, queries with $st(q) < 0.5$ make up 47% of the test set. Routing only these queries to the stability-conditioned 18-source configuration and sending the rest to Hybrid gives Coverage@20 of .582, compared with .588 for always using that configuration and .561 for Hybrid. This reduces retrieval calls from 18 to 9.5 per query, a 47% reduction, for a 0.6 pp Coverage@20 loss relative to always using the full configuration.

5.5 Downstream Effects and Practical Cost

Cost. Stability-conditioned consensus uses 18 retrieval calls per query, compared with 2 for Hybrid. On HotpotQA/MiniLM, retrieval plus consensus takes about 21 ms per query on an H100, compared with about 7 ms for Hybrid. Generating five paraphrases in one Qwen3-32B call takes about 1.7 s in the measured online setup, although the paraphrases can be cached when queries are known in advance. The routing result above reduces retrieval cost when the stability-conditioned consensus configuration is not needed.

Downstream answer quality. Table 7 evaluates Qwen3-32B with the top five retrieved evidence items. Stability-conditioned consensus improves the downstream score in all nine settings by 0.1–2.3 pp, with 7/9 differences significant at $p < 0.05$. The gains are modest in the near-ceiling FEVER settings. A separate GPT-5.4 grounding audit over 1,800 judgments labels 74.2% of stability-conditioned-consensus answers as fully

Method	Calls*	HotpotQA	FEVER	SQuAD 2.0
<i>Single-source retrieval</i>				
Dense (MiniLM)	1	.372	.887	.896
BM25	1	.498	.933	.944
SPLADE (Formal et al., 2021)	1	.610	.952	.966
<i>Two-source fusion</i>				
Hybrid RRF (Dense+BM25)	2	.561	.945	.965
SPLADE+Dense RRF	2	.588	.951	.962
<i>Query / multi-source retrieval</i>				
ColBERTv2 (Santhanam et al., 2022)	1	.570	.938	.981
HyDE (Gao et al., 2023)	1 [†]	.480	.941	.893
RAG-Fusion (Rackauckas, 2024)	6 [†]	.385	.904	.898
All-source RRF	18 [†]	.553	.947	.970
Uniform consensus	18 [†]	.570	.949	.974
Stab.-cond. consensus (BM25)	18 [†]	.588	.951	.975
Stab.-cond. consensus (SPLADE)	18 [†]	.617	.953	.972
18-source pool ceiling	18 [†]	.725	.967	.992

Table 6: Coverage@20 on MiniLM unless noted. *Calls denotes retriever calls per query. [†]The method also uses one Qwen3-32B call for paraphrase or hypothetical-document generation.

	Hybrid	Stability-cond. consensus	Δ
<i>HotpotQA (507K), F1</i>			
MiniLM	50.1	51.5	+1.4*
Qwen3-0.6B	52.2	53.6	+1.4*
Qwen3-8B	55.8	57.2	+1.4*
<i>FEVER (14K), Accuracy</i>			
MiniLM	90.4	90.7	+0.3
Qwen3-0.6B	90.5	90.6	+0.1
Qwen3-8B	90.5	91.0	+0.5*
<i>SQuAD 2.0 (20K), F1</i>			
MiniLM	58.9	61.2	+2.3*
Qwen3-0.6B	60.5	61.3	+0.8*
Qwen3-8B	61.4	62.0	+0.6*

Table 7: End-to-end RAG quality with Qwen3-32B and top-5 evidence. * $p < 0.05$ by paired bootstrap.

supported by the retrieved evidence, compared with 72.8% for Hybrid.

6 Conclusion

Meaning-preserving paraphrases can cause large changes in retrieved evidence, and this instability remains across datasets, model sizes, and encoder families. Additional characterization shows that lexical distance is only modestly associated with retrieval change, while substantial instability remains even for fully faithful and lexically conservative reformulations. Multi-source consensus improves retrieval, and fixed source-type weighting can add further gains in harder settings. However, making those weights query-adaptive with $st(q)$ changes Coverage@20 only slightly and can be

negative when retrieval is already stable. Retrieval stability is more useful as a label-free routing signal for deciding when the extra cost of consensus is worthwhile. Downstream gains are also modest in settings where retrieval is already near ceiling. These results suggest that reliability under linguistic variation should be measured separately from average retrieval accuracy.

Limitations

Our experiments use three English, Wikipedia-based benchmarks and constructed retrieval corpora. They cover multi-hop QA, fact verification, and reading comprehension, but they do not show whether the same patterns hold for web-scale retrieval, domain-specific corpora, multilingual settings, or proprietary RAG systems. E5-large-v2 and BGE-large-en-v1.5 broaden the encoder-family check, but the full mitigation experiments are run only with MiniLM and the two Qwen3-Embedding models.

We use LLM-generated paraphrases as a controlled probe of wording variation. NLI checks, an LLM fidelity audit, high-fidelity-only analysis, and a Llama-3.1-8B paraphraser check reduce the risk that the result is caused only by poor paraphrases, but they do not replace naturally occurring user reformulations. Studying human-written query variants is an important next step. Our analysis of linguistic change is correlational: content-word distance, name or number changes, and fidelity-audit categories characterize variation

but do not directly measure ambiguity or establish what causes the instability.

The Qwen3-Embedding experiments use raw query and document text without a query-side task instruction. The official Qwen3-Embedding usage guidance recommends query instructions for retrieval, so our setting should be read as a controlled prompt-neutral comparison rather than the recommended deployment setup. We do not test whether adding such instructions changes the measured instability.

The downstream evaluation uses one generator family, Qwen3-32B, and the grounding analysis uses an LLM judge rather than human annotation. Broader generator and human evaluation would strengthen the downstream conclusions. Finally, stability-conditioned consensus uses more retrieval calls than Hybrid, and its gains are small in several strong or near-ceiling settings. The routing result reduces this cost in one representative setting, but broader deployment studies are needed to determine when the added computation is worthwhile.

References

- Beijing Academy of Artificial Intelligence. 2024. [BGE-reranker-v2-gemma](#). Model card.
- David Carmel and Elad Yom-Tov. 2010. *Estimating the Query Difficulty for Information Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Gordon V. Cormack, Charles L.A. Clarke, and Stefan Buettcher. 2009. [Reciprocal rank fusion outperforms condorcet and individual rank learning methods](#). In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, Boston, MA, USA. Association for Computing Machinery.
- Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. [SPLADE: Sparse lexical and expansion model for first stage ranking](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292, Virtual Event, Canada. Association for Computing Machinery.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. [Precise zero-shot dense retrieval without relevance labels](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 540 others. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Faisal Hamman, Chenyang Zhu, Anoop Kumar, Xujun Peng, Sanghamitra Dutta, Daben Liu, and Alf Samuel. 2025. [Improving consistency in retrieval-augmented systems with group similarity rewards](#). In *NeurIPS 2025 Workshop on Reliable ML from Unreliable Data*.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). In *The Eleventh International Conference on Learning Representations*.
- Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with PagedAttention](#). In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, Koblenz, Germany. Association for Computing Machinery.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. [Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and*

- Development in Information Retrieval*, pages 2356–2362, Virtual Event, Canada. Association for Computing Machinery.
- Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. [Query rewriting in retrieval-augmented large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, Singapore. Association for Computational Linguistics.
- Shengyu Mao, Yong Jiang, Boli Chen, Xiao Li, Peng Wang, Xinyu Wang, Pengjun Xie, Fei Huang, Huanjun Chen, and Ningyu Zhang. 2024. [RaFe: Ranking feedback improves query rewriting for RAG](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 884–901, Miami, Florida, USA. Association for Computational Linguistics.
- Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. [MTEB: Massive text embedding benchmark](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.
- Eviatar Nachshoni, Arie Cattan, Shmuel Amar, Ori Shapira, and Ido Dagan. 2025. [Consensus or conflict? fine-grained evaluation of conflicting answers in question-answering](#). In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertainNLP 2025)*, pages 138–159, Suzhou, China. Association for Computational Linguistics.
- OpenAI. 2026. [Introducing GPT-5.4](#).
- Zackary Rackauckas. 2024. [RAG-Fusion: A new take on retrieval-augmented generation](#). *arXiv preprint arXiv:2402.03367*.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don’t know: Unanswerable questions for SQuAD](#). *arXiv preprint arXiv:1806.03822*.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: BM25 and beyond](#). *Foundations and Trends in Information Retrieval*, 3(4):333–389.
- Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. [ColBERTv2: Effective and efficient retrieval via lightweight late interaction](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.
- Arnab Sharma, Daniel Vollmers, and Axel-Cyrille Ngonga Ngomo. 2025. [Calibrating language models for neural ranking under noisy supervision with relaxed labels](#). In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertainNLP 2025)*, pages 259–272, Suzhou, China. Association for Computational Linguistics.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. [BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
- Sergey Troshin, Irina Saporina, Antske Fokkens, and Vlad Niculae. 2025. [Asking a language model for diverse responses](#). In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertainNLP 2025)*, pages 66–72, Suzhou, China. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. [Text embeddings by weakly-supervised contrastive pre-training](#). *arXiv preprint arXiv:2212.03533*.
- Liang Wang, Nan Yang, and Furu Wei. 2023. [Query2doc: Query expansion with large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423, Singapore. Association for Computational Linguistics.
- William Webber, Alistair Moffat, and Justin Zobel. 2010. [A similarity measure for indefinite rankings](#). *ACM Transactions on Information Systems*, 28(4):20:1–20:38.
- Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. [Certifiably robust RAG against retrieval corruption](#). *arXiv preprint arXiv:2405.15556*.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2023. [C-pack: Packed resources for general chinese embeddings](#). *arXiv preprint arXiv:2309.07597*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Qwen3 Embedding: Advancing text embedding and reranking through foundation models](#). *arXiv preprint arXiv:2506.05176*.

A Implementation Details and Reproducibility

A.1 Consensus Details

Selected source weights are shown in Table 8.

$st(q)$	w_d	w_b	w_s	w_{pd}	w_{pb}	w_{ps}
0.0 (unstable)	0.30	1.50	0.20	0.12	0.60	0.08
0.5 (threshold)	0.55	1.00	0.20	0.22	0.40	0.08
1.0 (stable)	0.80	0.50	0.20	0.32	0.20	0.08

Table 8: Selected source weights at three $st(q)$ values. $d/b/s$ denote dense, BM25, and sentence-mapped sources; p denotes the paraphrase-discounted counterpart ($\gamma = 0.4$).

Algorithm 1 SCER aggregation for a single query.

Require: Query q ; m paraphrases $\mathcal{P}(q) = \{q_1, \dots, q_m\}$; retrieval depth K ; output size k_{out} ; weight params $(a, b, c, d, w_s, \gamma)$.

- 1: For each $q_i \in \{q\} \cup \mathcal{P}(q)$, retrieve top- K from 3 sources: dense paragraph $L_{q_i}^d$, BM25 paragraph $L_{q_i}^b$, and sentence-mapped $L_{q_i}^s$.
- 2: Compute $st(q) \leftarrow \frac{1}{m} \sum_{i=1}^m J(L_q^d, L_{q_i}^d)$ $\triangleright$ Eq. 1: dense-only, top- K
- 3: Set query-conditional weights $w_d \leftarrow a + b st(q)$, $w_b \leftarrow c - d st(q)$, $w_s \leftarrow w_s$.
- 4: Initialize $S[e] \leftarrow 0$ for all documents e .
- 5: **for** each query variant q_i and retriever type $t \in \{d, b, s\}$ **do**
- 6: $w \leftarrow w_t$ if $q_i = q$ else $\gamma \cdot w_t \triangleright$ paraphrase discount
- 7: **for** each e at rank r in $L_{q_i}^t$, $r = 1, \dots, K$ **do**
- 8: $S[e] \leftarrow S[e] + w/r$
- 9: **end for**
- 10: **end for**
- 11: Return top- k_{out} documents by $S[e]$, with ties broken by the order in which documents were first inserted into S during the loop above (stable sort over insertion order).

A.2 Retrieval Implementation

Dense retrieval uses exact PyTorch inner-product search over L2-normalized embeddings. FAISS CPU 1.13 is used only to store embedding caches. BM25 uses Pyserini 1.2 (Lin et al., 2021) with default parameters ($k_1 = 0.9$, $b = 0.4$) and the default English analyzer. MiniLM uses sentence-transformers 3.4. Qwen3-Embedding-0.6B/8B use the official checkpoints with vLLM 0.16 (Kwon et al., 2023) and last-token pooling, followed by L2 normalization. We pass raw query and document text without task-specific query instructions; this choice is discussed in the Limitations section.

The Wikipedia-derived corpora may overlap with the pretraining data of the embedding and generation models, and we do not control for this. Qwen3-32B generates the main paraphrases, while Qwen3-Embedding checkpoints are used for retrieval. A separate Llama-3.1-8B paraphraser gives the same overall instability pattern. We do not report the Llama-paraphrase sensitivity run with Qwen3-Embedding-8B because that auxil-

iary re-encoding run used mean pooling rather than the last-token pooling used in the main pipeline.

Main encoding batch sizes are 1,024 for MiniLM and 256 for Qwen3-Embedding-0.6B/8B. Sentence-level encoding is sharded at 2,048, dense top- K search batches queries in groups of 5,000, and paraphrase generation uses a vLLM batch of 256. LLM components run with Qwen3-32B-FP8 on one NVIDIA H100 80GB GPU.

A.3 Paraphrase Generation and Quality Checks

For each query q , Qwen3-32B-FP8 generates five paraphrases in one call at temperature 0.7 and a 300-token limit. The prompt is:

```
Generate exactly 5
paraphrases of this
question. Each paraphrase
must:
- Ask exactly the same
thing
- Use different words and
sentence structure
- Be a complete, natural
question.
Question: {q}. Output
ONLY 5 paraphrases, one
per line, numbered 1-5.
No other text.
```

Outputs are parsed from numbered lines, with a fallback that does not require a final question mark. On HotpotQA/MiniLM, using $m = \{1, 3, 5, 10\}$ paraphrases gives Coverage@20 of $\{.585, .585, .588, .586\}$, so we use $m = 5$.

We validate paraphrases in two ways. First, DeBERTa-v3 MNLI over all query-paraphrase pairs gives bidirectional entailment rates of 86.7%, 91.6%, and 88.3% on HotpotQA, FEVER, and SQuAD 2.0, with contradiction below 2.4%. Second, a GPT-5.4 audit of 1,500 stratified paraphrases gives mean fidelity scores of 4.56, 4.57, and 4.54 out of 5, with 89.8%, 89.2%, and 90.4% rated at least 4. The main low-fidelity cases are subject changes (56/1500), scope changes (26/1500), and factual errors (20/1500).

We do not filter paraphrases before retrieval. On the audited high-fidelity subset, top-20 Jaccard distance remains 41–49% on HotpotQA, 38–47% on FEVER, and 31–40% on SQuAD 2.0

across the three main encoders. Low-fidelity paraphrases increase instability further on HotpotQA and FEVER, but they are not needed for the main effect. The Llama-3.1-8B paraphraser sensitivity results are reported in the main paper.

A.4 Paraphrase Change Analysis

We analyze all five Qwen3-32B paraphrases for every query, giving 37,025, 32,530, and 29,640 original-paraphrase pairs per encoder for HotpotQA, FEVER, and SQuAD 2.0, respectively. Content-word overlap is Jaccard overlap after lowercasing, tokenizing runs of ASCII letters and digits, and removing the 198-word NLTK English stopword list. Retrieval change is one minus Jaccard overlap between the corresponding dense top-20 sets. We also report *gold-evidence miss*: among pairs for which the original top-20 contains all gold evidence, the rate at which the paraphrase top-20 no longer does so.

Because the five paraphrases of a query are not independent, 95% intervals use 1,000 bootstrap resamples of queries (seed 0). A simple name-or-number flag records whether an original containing a capitalized token after the first word or a token containing a digit loses at least one such token in the paraphrase. To separate this flag from overall lexical change, we fit least-squares models of each outcome on content-word distance and the flag with one intercept per query. These analyses are correlational rather than causal. Repeating the lexical analysis with the broader 318-word scikit-learn stopword list changes no conclusion; the largest change in within-query rank correlation is 0.012.

Detailed results are reported in Tables 9–12.

A.5 Downstream and Baseline Details

Baseline implementations. RAG-Fusion uses the same five generated paraphrases as SCER, dense retrieval for each query variant, and RRF with $k_0 = 60$. All-source RRF applies the same RRF rule to all 18 source lists. Uniform SCER uses the same 18 lists and paraphrase discount $\gamma = 0.4$, with all source-type weights set to 1.

End-to-end generation. For HotpotQA and SQuAD 2.0, the generator is asked for a short answer based on the provided evidence. For FEVER, it predicts exactly SUPPORTS or REFUTES. The top five retrieved paragraphs are passed in rank order, each truncated to 500 characters. Qwen3-

Dataset	Encoder	Rank correlation all pairs	Rank correlation within query	Top-20 change most / least reworded	Gold-evidence miss most / least reworded	Gold-evidence miss gap (pp)
HotpotQA	MiniLM	.28 [.27,.29]	.26 [.24,.27]	55.6% / 43.4%	17.7% / 11.2%	+6.4 [+4.6, +8.4]
	Qwen3-0.6B	.20 [.19,.21]	.18 [.16,.19]	54.0% / 45.7%	14.5% / 8.8%	+5.7 [+4.4, +7.3]
	Qwen3-8B	.19 [.17,.20]	.17 [.16,.18]	45.8% / 38.4%	10.5% / 5.1%	+5.4 [+4.4, +6.5]
FEVER	MiniLM	.28 [.27,.30]	.26 [.25,.27]	53.3% / 41.1%	1.7% / 1.0%	+0.7 [+0.3, +1.1]
	Qwen3-0.6B	.15 [.14,.17]	.15 [.13,.16]	48.6% / 42.6%	0.6% / 0.4%	+0.2 [-0.1, +0.4]
	Qwen3-8B	.20 [.18,.21]	.14 [.13,.15]	42.4% / 34.7%	0.6% / 0.1%	+0.5 [+0.3, +0.7]
SQuAD 2.0	MiniLM	.32 [.31,.34]	.28 [.27,.30]	48.4% / 34.4%	5.5% / 2.1%	+3.4 [+2.7, +4.2]
	Qwen3-0.6B	.23 [.21,.24]	.22 [.21,.23]	42.5% / 33.0%	3.7% / 1.4%	+2.3 [+1.8, +2.9]
	Qwen3-8B	.23 [.21,.24]	.20 [.19,.21]	36.8% / 27.9%	2.9% / 0.9%	+2.0 [+1.5, +2.6]

Table 9: Lexical change and retrieval change. Pairs are divided into thirds by content-word overlap within each dataset; “most/least reworded” compares the lowest- and highest-overlap thirds. Brackets are 95% query-bootstrap intervals.

Dataset	Encoder	Pairs	Top-20 change	Gold-evidence miss
HotpotQA	MiniLM	669	35.5% [34.3,36.8]	9.3%
	Qwen3-0.6B	669	44.9% [43.5,46.4]	8.6%
	Qwen3-8B	669	37.7% [36.4,39.0]	4.7%
SQuAD 2.0	MiniLM	189	24.3% [22.0,26.7]	1.8%
	Qwen3-0.6B	189	30.5% [28.2,32.9]	1.9%
	Qwen3-8B	189	25.0% [22.6,27.4]	1.1%

Table 10: Retrieval change for paraphrases that preserve exactly the same set of content words as the original. FEVER has only two such pairs and is omitted.

32B-FP8 runs at temperature 0 with a 50-token output limit. QA outputs are evaluated with normalized exact match and token F1; FEVER uses accuracy.

Grounding audit. The GPT-5.4 grounding audit contains 1,800 judgments: 100 sampled queries per setting, nine settings, and both Hybrid and SCER answers. Each answer is labeled SUPPORTED, PARTIALLY SUPPORTED, NOT SUPPORTED, or CONTRADICTED using its retrieved top-five evidence. Across all settings, 74.2% of SCER answers and 72.8% of Hybrid answers are labeled SUPPORTED. Because Qwen3-32B is used for both paraphrase generation and downstream answer generation, broader generator-family evaluation remains future work.

A.6 Stability-Conditioned Weighting Details

For completeness, the tested query-adaptive variant uses:

$$w_{\text{dense}}(q) = a + b \text{ st}(q), \quad (4)$$

$$w_{\text{BM25}}(q) = c - d \text{ st}(q), \quad (5)$$

$$w_{\text{smapped}} = w_s, \quad (6)$$

$$w_{\text{para-*}}(q) = \gamma w_*(q), \quad (7)$$

$$* \in \{\text{dense}, \text{BM25}, \text{smapped}\}.$$

The selected parameters are $(a, b, c, d, w_s, \gamma) = (0.3, 0.5, 1.5, 1.0, 0.2, 0.4)$.

B Extended Results

B.1 Instability and Stability Analysis

Tables 13 and 14, Figure 3, and Table 15 provide the detailed stability analyses summarized in the main paper.

B.2 Consensus Ablations and Robustness

Tables 16–22 report the extended metrics, pool analysis, ablations, split robustness, rank-discount checks, and significance tests.

B.3 Reranking Analysis

Tables 23–25 report candidate-pool recall and two reranking experiments.

To check that the pattern in Table 24 is not specific to Qwen3-Reranker-8B, we re-ran the same protocol with BGE-reranker-v2-gemma (Chen et al., 2024; Beijing Academy of Artificial Intelligence, 2024) (2B parameters, Gemma-2B backbone scored via the “Yes” token logit). Results in Table 25 show the same three-regime shape with smaller absolute deltas: SCER’s pool *helps* the reranker on HotpotQA (low base recall, +1.4 to +3.2 pp), is essentially neutral on FEVER, and modestly hurts on SQuAD 2.0 (−1.1 to −1.2 pp vs. −1.7 to −2.5 pp with Qwen3-Reranker-8B). The shared pattern across two rerankers suggests an operational reading: SCER and reranking occupy adjacent points on the recall–quality curve. Stacking them often helps on HotpotQA, especially in lower-recall settings, but can be neutral or negative when the base retriever is already strong (e.g., Qwen3-Reranker-8B on HotpotQA/Qwen3-8B: S+R .756 vs. H+R .765, −0.9 pp). On near-saturated pools (SQuAD 2.0, R@20 ≥ .97),

Dataset	Encoder	Top-20 change changed / kept	Gold-evidence miss changed / kept	Within-query top-20 difference (pp)	Controlled additional top-20 change (pp)	Controlled additional gold-evidence miss (pp)
HotpotQA	MiniLM	55.9% / 48.9%	19.2% / 13.3%	+6.9 [+6.0, +7.7]	+2.4 [+1.6, +3.2]	+4.4 [+1.7, +7.3]
	Qwen3-0.6B	53.6% / 49.3%	15.6% / 10.4%	+5.5 [+4.8, +6.3]	+3.0 [+2.3, +3.7]	+3.7 [+1.5, +5.9]
	Qwen3-8B	44.2% / 41.5%	11.9% / 6.8%	+4.9 [+4.1, +5.7]	+2.3 [+1.6, +3.1]	+4.4 [+2.9, +6.0]
FEVER	MiniLM	49.3% / 47.8%	1.9% / 1.4%	+3.6 [+2.8, +4.5]	−1.7 [−2.6, −1.0]	−0.1 [−0.9, +0.7]
	Qwen3-0.6B	46.4% / 45.6%	0.6% / 0.4%	+3.8 [+3.1, +4.5]	+1.3 [+0.7, +2.0]	+0.3 [−0.2, +0.8]
	Qwen3-8B	39.9% / 38.4%	0.6% / 0.3%	+3.5 [+2.8, +4.3]	+1.3 [+0.7, +2.0]	+0.3 [−0.1, +0.8]
SQuAD 2.0	MiniLM	45.1% / 39.0%	3.8% / 2.3%	+6.2 [+5.2, +7.1]	+0.8 [−0.1, +1.8]	+1.5 [+0.1, +3.1]
	Qwen3-0.6B	40.5% / 35.4%	3.0% / 1.7%	+4.9 [+4.0, +5.8]	+1.5 [+0.6, +2.4]	+0.5 [−0.3, +1.6]
	Qwen3-8B	33.5% / 30.2%	1.4% / 1.1%	+3.8 [+3.0, +4.7]	+0.7 [−0.2, +1.5]	+0.1 [−0.5, +0.7]

Table 11: Name-or-number changes. The controlled columns use content-word distance plus a name-or-number-change indicator with one intercept per query. Brackets are 95% query-bootstrap intervals.

Audit label	Pairs	Top-20 change (%)			Gold-evidence miss (%)		
		MiniLM	0.6B	8B	MiniLM	0.6B	8B
Fully faithful (score 5)	1,100	44.8	43.5	36.2	3.5	4.0	2.7
Faithful (score 4)	247	48.3	45.0	39.3	6.5	2.5	3.2
No failure label, score ≤ 3	50	44.5	40.7	34.2	0.0	2.4	2.2
Subject change	56	55.8	56.4	50.2	14.3	15.4	10.6
Scope change	26	45.9	44.1	40.8	11.1	5.0	0.0
Factual error	20	42.6	40.6	36.6	11.1	20.0	23.5
Entity substitution	1	40.0	75.0	57.1	0.0	0.0	0.0

Table 12: Retrieval change by GPT-5.4 paraphrase-fidelity audit category. Fully faithful paraphrases still change 36–45% of the top-20. Small flagged categories should be interpreted cautiously.

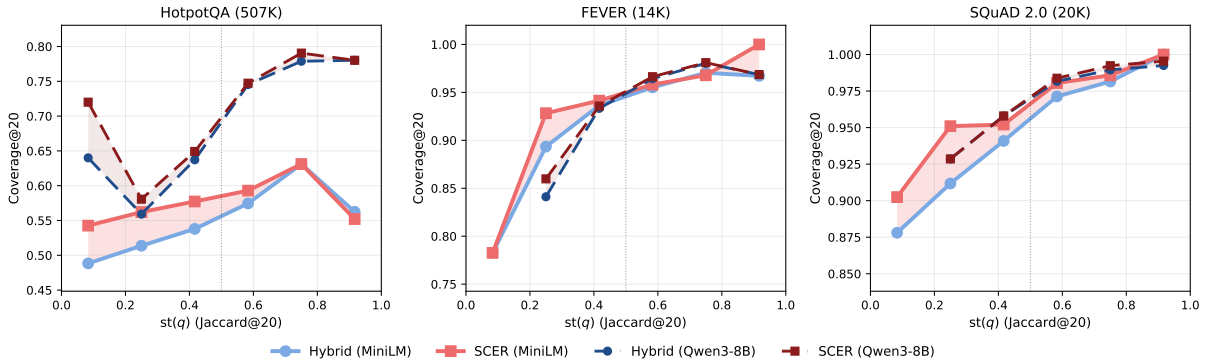


Figure 3: Absolute Coverage@20 by retrieval-stability bin. The gap between SCER and Hybrid is generally larger in lower-stability bins.

Benchmark	Metric	MiniLM	Qwen3-0.6B	Qwen3-8B
HotpotQA (507K)	$st(q)$	0.50 ± 0.16	0.50 ± 0.15	0.58 ± 0.14
	Replace rate	0.355	0.354	0.282
	RBO ($p=0.9$)	0.591	0.605	0.668
FEVER (14K)	$st(q)$	0.52 ± 0.14	0.54 ± 0.13	0.61 ± 0.13
	Replace rate	0.333	0.310	0.252
	RBO ($p=0.9$)	0.619	0.653	0.692
SQuAD 2.0 (20K)	$st(q)$	0.59 ± 0.16	0.63 ± 0.16	0.68 ± 0.14
	Replace rate	0.279	0.248	0.205
	RBO ($p=0.9$)	0.622	0.653	0.696

Table 13: Detailed instability metrics for the three main encoders. $st(q)$ is Jaccard@20 overlap (mean $\pm$ SD); Replace rate is the mean fraction of top-20 documents replaced; RBO is Rank-Biased Overlap (Webber et al., 2010) with $p = 0.9$.

reranker reordering noise dominates and SCER alone is the better choice.

B.4 Weight and Parameter Sensitivity

Where the method fails. The full tuned weighting schedule can underperform uniform consensus in already-stable settings: SCER vs. uniform-weight SCER is -0.04 to -0.25 pp Cov@20 on the three Qwen3-8B cells (HotpotQA/FEVER/SQuAD 2.0). This comparison includes both fixed source-type weights and the query-adaptive $st(q)$ component, so the negative difference should not be attributed to stability adaptation alone. The isolated $st(q)$ -adaptive component changes Coverage@20 by at most ± 0.3 pp per setting (Table 18).

Leave-one-setting-out weight sensitivity. We ran a coarser 729-config grid search nine times,

			MiniLM		Qwen3-0.6B		Qwen3-8B	
Stability		n	Hybrid	Δ	Hybrid	Δ	Hybrid	Δ
HQA	Unstable	2.8K	.529	+4.3	.559	+2.4	.624	+1.5
	Stable	3.1K	.590	+1.2	.675	+1.6	.759	+0.5
FEV	Unstable	2.3K	.926	+1.1	.923	+0.6	.923	+0.4
	Stable	2.9K	.960	+0.2	.968	+0.0	.971	+0.1
SQD	Unstable	1.3K	.932	+1.8	.949	+1.3	.952	+0.0
	Stable	3.4K	.978	+0.6	.982	+0.3	.987	+0.3

Table 14: Coverage@20 stratified by stability (rotated labels: HQA = HotpotQA, FEV = FEVER, SQD = SQuAD 2.0; Δ = SCER – Hybrid, pp). Unstable: $\text{st}(q) < 0.5$ (midpoint of Jaccard $[0, 1]$, not a quantile); SQuAD 2.0’s 28/72% split reflects that most queries are above the midpoint. Continuous version: Fig. 3.

Setting	$r(\text{st}, d)$	$r(\text{st}, b)$	$r(\text{st}, \Delta_{\text{SCER}-H})$	$r(d, b)$
HotpotQA/MiniLM	+.129	+.017	−.063	+.359
HotpotQA/Qwen3-0.6B	+.227	+.097	−.004	+.393
HotpotQA/Qwen3-8B	+.160	+.108	+.005	+.381
FEVER/MiniLM	+.169	+.060	−.088	+.453
FEVER/Qwen3-0.6B	+.112	+.049	−.078	+.635
FEVER/Qwen3-8B	+.111	+.067	+.003	+.545
SQuAD 2.0/MiniLM	+.215	+.081	−.041	+.265
SQuAD 2.0/Qwen3-0.6B	+.218	+.092	−.048	+.258
SQuAD 2.0/Qwen3-8B	+.144	+.095	+.045	+.193

Table 15: Per-query Pearson correlations (test set, capped at 2,000 queries/setting). d, b, h, s denote dense, BM25, Hybrid, SCER per-query Cov@20. $\text{st}(q)$ correlates more strongly with dense success (+.11–+.23) than with BM25 success (+.02–+.11): it is a dense-reliability signal, not a BM25-quality predictor.

Method	Cov@5	Cov@20	MRR	nDCG@20
Dense	.181	.372	.499	.460
BM25	.288	.498	.612	.578
Hybrid RRF	.397	.561	.747	.660
All-source RRF	.443	.553	.769	.690
Uniform SCER	.467	.570	.768	.693
SCER	.487	.588	.769	.695

Table 16: Extended retrieval metrics on HotpotQA/MiniLM ($n = 5,924$). Rank-sensitive MRR/nDCG@20 separate the 18-source family from Hybrid more sharply than Cov@20.

Benchmark	Recall@N (mean fraction)						Cov@ U (strict)
	10	20	40	60	100	U	
HotpotQA	.705	.765	.810	.832	.851	.860	.725
FEVER	.957	.968	.975	.978	.980	.982	.967
SQuAD 2.0	.939	.970	.983	.987	.990	.992	.992

Table 17: 18-source pool depth (MiniLM). FEVER/SQuAD 2.0 nearly saturate by $N = 20$; HotpotQA continues to gain with deeper N , and the 14 pp Recall–Cov gap tracks multi-hop queries with only one of two gold paragraphs in the pool.

holding out one (benchmark, embedder) setting per fold and tuning on the remaining 8. Per-fold tuned weights match the paper’s joint-tuned

Configuration	HotpotQA			FEVER			SQuAD 2.0		
	MiniLM	0.6B	8B	MiniLM	0.6B	8B	MiniLM	0.6B	8B
SCER (full)	.588	.637	.729	.951	.954	.964	.975	.980	.985
– paraphrase queries	.578	.624	.716	.948	.953	.963	.972	.977	.985
– sentence channel	.578	.627	.721	.950	.954	.964	.970	.976	.981
– BM25 channel	.435	.555	.723	.928	.943	.960	.940	.956	.976
– $\text{st}(q)$ -adaptive weights	.588	.634	.727	.950	.955	.964	.974	.980	.985
Uniform SCER ($\gamma = 0.4$)	.570	.628	.732	.949	.953	.964	.974	.979	.986
All-source RRF	.553	.618	.728	.947	.952	.964	.970	.976	.984

Table 18: Full component ablation (Coverage@20). Rows remove one component from full SCER or replace its aggregation rule. This table keeps the detailed channel-level results that are summarized in the main paper.

Split seed	Mean gain (pp)	Positive	Retuned gain (pp)
42	0.91	9/9	0.91
0	0.96	9/9	0.87
1	0.98	9/9	1.00
2	0.89	9/9	0.87
3	0.90	9/9	0.89
Mean	0.93	9/9	0.91

Table 19: Robustness of the SCER Coverage@20 gain over Hybrid across five 20/80 calibration/test splits. Mean gain uses the paper weights; Retuned gain re-selects weights on each split.

Comparison	Mean $\pm$ SD (pp)
Uniform SCER – Hybrid	+0.60 $\pm$ 0.03
Paper weighting – Uniform SCER	+0.33 $\pm$ 0.02
Reversed weighting – Uniform SCER	+0.02 $\pm$ 0.02
Paper weighting – Reversed weighting	+0.31 $\pm$ 0.02

Table 20: Weighting-direction check over the same five splits. The selected weighting direction is better than the reversed direction; Table 5 separately isolates the query-adaptive $st(q)$ component.

Setting	1/r (paper)	RRF $1/(60+r)$	Borda $K-r+1$	Unif 1
HotpotQA/MiniLM	.588	.576	.587	.569
HotpotQA/0.6B	.637	.626	.637	.619
HotpotQA/8B	.729	.721	.730	.715
FEVER/MiniLM	.951	.951	.950	.951
FEVER/0.6B	.954	.956	.954	.956
FEVER/8B	.964	.966	.965	.965
SQuAD 2.0/MiniLM	.975	.974	.975	.973
SQuAD 2.0/0.6B	.980	.978	.979	.976
SQuAD 2.0/8B	.985	.984	.986	.983

Table 21: Coverage@20 under four rank-discount choices (same SCER 18-source pool and source weights; only the per-rank discount changes). $1/r$ is the paper default.

Setting	Hybrid	SCER	Δ (pp)	p	$\pm$ CI (pp)
HotpotQA/MiniLM	.561	.588	+2.68	$< 10^{-4} *$	± 1.25
HotpotQA/0.6B	.618	.637	+1.98	$< 10^{-4} *$	± 1.22
HotpotQA/8B	.721	.729	+0.79	.0015*	± 1.13
FEVER/MiniLM	.945	.951	+0.61	$< 10^{-4} *$	± 0.59
FEVER/0.6B	.952	.954	+0.23	.0258	± 0.57
FEVER/8B	.962	.964	+0.19	.0229	± 0.52
SQuAD 2.0/MiniLM	.965	.975	+0.97	$< 10^{-4} *$	± 0.44
SQuAD 2.0/0.6B	.974	.980	+0.53	.0001*	± 0.39
SQuAD 2.0/8B	.983	.985	+0.23	.0152	± 0.35

Table 22: Per-setting Cov@20 for Hybrid and SCER with exact paired-bootstrap p -value ($n = 10,000$ resamples) and SCER 95% CI half-width. * survives Bonferroni at $\alpha = 0.05/9 = .0056$. The $\pm$ CI column is the marginal SCER CI, not the paired-difference CI; paired-bootstrap p -values can therefore survive even where the marginal CI touches zero.

weights within ± 0.5 pp Cov@20 on every fold (mean diff -0.07 pp). Slope b and discount γ are identical in all folds; base weights a , c , d vary modestly ($sd \leq 0.34$). The combination of small per-fold swings, zero variance on b and γ , and only modest sd on $a/c/d$ indicates a *broad plateau* rather than a narrow optimum: the linear weighting form is well-conditioned and the tuning step behaves like a calibration over a flat region, consistent with the small headline $+0.30$ pp gain over uniform weights.

Benchmark	Embedder	Hybrid R@50	SCER R@50
HotpotQA	MiniLM	.817	.845
	Qwen3-0.6B	.852	.866
	Qwen3-8B	.900	.909
FEVER	MiniLM	.978	.980
	Qwen3-0.6B	.980	.982
	Qwen3-8B	.988	.989
SQuAD 2.0	MiniLM	.978	.986
	Qwen3-0.6B	.984	.993
	Qwen3-8B	.992	.995

Table 23: Recall@50 of the candidate pool fed to the cross-encoder reranker (1,000-query subsample). SCER’s pool has higher recall than Hybrid in all 9 cells, so reranker failures are reordering errors, not missing candidates.

Benchmark	Embedder	Hybrid	SCER	H+R	S+R	Δ
HotpotQA	MiniLM	.576	.605	.630	.645	+1.5
	Qwen3-0.6B	.652	.672	.686	.699	+1.3
	Qwen3-8B	.746	.759	.765	.756	-0.9
FEVER	MiniLM	.951	.957	.930	.911	-1.9
	Qwen3-0.6B	.959	.960	.926	.906	-2.0
	Qwen3-8B	.971	.969	.932	.909	-2.3
SQuAD 2.0	MiniLM	.968	.978	.946	.926	-2.0
	Qwen3-0.6B	.980	.982	.949	.932	-1.7
	Qwen3-8B	.984	.986	.958	.933	-2.5

Table 24: Cross-encoder reranking with Qwen3-Reranker-8B (1,000-query subsample, top-50 reranked to top-20). Input-pool Recall@50 in Table 23; second-reranker replication in Table 25.

Benchmark	Embedder	Hybrid	SCER	H+R	S+R	Δ
HotpotQA	MiniLM	.579	.592	.634	.664	+3.0
	Qwen3-0.6B	.621	.641	.669	.701	+3.2
	Qwen3-8B	.728	.736	.756	.770	+1.4
FEVER	MiniLM	.951	.959	.963	.964	$+0.1$
	Qwen3-0.6B	.961	.965	.961	.960	-0.1
	Qwen3-8B	.969	.971	.967	.966	-0.1
SQuAD 2.0	MiniLM	.961	.972	.947	.935	-1.2
	Qwen3-0.6B	.971	.979	.953	.942	-1.1
	Qwen3-8B	.978	.983	.957	.945	-1.2

Table 25: Reranker robustness: BGE-reranker-v2-gemma (2B, Gemma-2B backbone, CausalLM “Yes”-token scoring), top-50 reranked to top-20. Same 1,000-query subsample as Table 24.

Parameter sensitivity and tuning. Cov@20 at $m = \{1, 3, 5, 10\}$ paraphrases is $\{.585, .585, .588, .586\}$ on HotpotQA/MiniLM, with no consistent gain beyond $m = 5$. The main grid search evaluates 14,000 configurations on the 20% calibration split and selects $(a, b, c, d, w_s, \gamma) = (0.3, 0.5, 1.5, 1.0, 0.2, 0.4)$.