

Dual-View Training for Instruction-Following Information Retrieval

Qingcheng Zeng^{1,2*}, Puxuan Yu², Aman Mehta², Fuheng Zhao², Rajhans Samdani²

¹Northwestern University ²Snowflake Inc.

Abstract

Instruction-following information retrieval (IF-IR) requires retrievers to satisfy both topical intent and instance-specific relevance constraints. We propose dual-view training, a data synthesis strategy based on polarity reversal. Given a query, an instruction-relevant document, and a topically related document excluded by that instruction, an LLM generates a complementary instruction under which the two documents exchange relevance labels. The resulting paired views directly penalize instruction-agnostic representations. At a matched 880k-example budget, replacing the non-instruction half of a Promptriever-style joint-data recipe with polarity-reversed examples improves FollowIR p -MRR from 5.27 to 8.30 (+57.5%) on a 305M-parameter encoder while increasing its instruction-conditioned Score from 20.85 to 21.38. The relative p -MRR gain repeats on a 560M-parameter backbone. The proposed recipe also improves InfoSearch nDCG@5 and shows no average degradation on four evaluated BEIR datasets with plain queries. A size-matched paraphrase control does not reproduce the instruction-sensitivity gains, supporting the conclusion that label reversal, rather than additional LLM-generated wording, provides the key supervision signal.

1 Introduction

Instruction-following information retrieval extends traditional semantic matching by requiring systems to adhere to both a query and explicit user-defined constraints that specify relevance criteria (Su et al., 2023; Weller et al., 2025a). For instance, a user might not only submit a query but also specify that relevant documents must discuss a particular aspect, be written in a certain style, or satisfy length requirements. This paradigm rigorously tests the capacity of dense retrievers to adapt their behav-

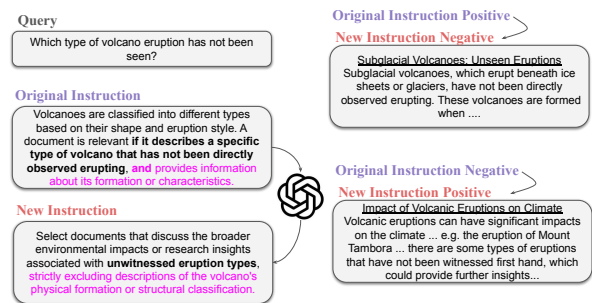


Figure 1: We synthesize new instructions that reverse the relevance polarity of existing documents, creating challenging samples that sharpen the retriever’s sensitivity to instructional nuances.

ior based on dynamic in-context directives, going beyond static notions of relevance.

Despite the growing number of instruction-aware retrievers, critical limitations persist. Weller et al. (2025a) conducted a systematic evaluation using human-annotated instructions that fundamentally alter relevance definitions. By quantifying sensitivity to instruction changes with the p -MRR metric, which measures whether a retriever ranks the preferred document higher when the instruction changes, their findings reveal that most current models fail to internalize detailed relevance criteria, relying instead on superficial query-document similarity and largely ignoring the specific constraints imposed by instructions.

To address this, Weller et al. (2025b) introduced a training paradigm centered on *instruction negatives*, documents that are semantically relevant to the query but become irrelevant once a specific instruction is applied. While their results demonstrate the effectiveness of instruction negatives over standard hard negatives, these negatives carry additional untapped potential: each one implicitly defines a complementary instruction under which it becomes the relevant document. In this paper, we exploit this observation by prompting an LLM

*Work done during internship at Snowflake Inc.

to synthesize such complementary instructions that reverse the relevance polarity of existing document pairs (Figure 1). The same documents are thus repurposed under two complementary views, compelling the retriever to attend to fine-grained instructional differences rather than surface-level query-document similarity.

We make three contributions. (1) We propose a simple polarity-reversal strategy that turns each instruction negative into a conditionally relevant document under a complementary instruction, creating two relevance views over the same query-document pair. (2) In the primary matched-budget comparison, our full dual-view recipe improves instruction sensitivity over a Promptriever-style joint-data control while preserving the evaluated aggregate instruction-conditioned and standard-retrieval effectiveness; the p -MRR gain repeats across two encoder backbones. (3) A size-matched paraphrase control, query-level significance tests, per-collection diagnostics, synthetic-data audits, and a prompted paraphrase evaluation characterize the source and reliability of the gains.

2 Methodology

As illustrated in Figure 1, our approach leverages LLMs to generate complementary instructions that invert the ground truth labels for a fixed set of documents. Formally, given a query q , a positive document D^+ , and an instruction negative document D^- under an original instruction I_{orig} , we prompt the LLM to synthesize a new instruction I_{new} . The generation is constrained such that I_{new} must be semantically coherent with q but sufficiently distinct from I_{orig} so that D^- becomes the relevant document (positive) and D^+ becomes an instruction negative. This creates a “dual-view” training scenario where the relevance of a document depends entirely on the specific constraints of the instruction, not on its text or the query alone.

During contrastive training, each query is paired with its positive document and a set of negatives. For data points augmented with our method, the training batch contains both the original and the polarity-reversed view. The model must therefore learn to assign high similarity to $(q \oplus I_{orig}, D^+)$ and $(q \oplus I_{new}, D^-)$ simultaneously, while pushing the reversed assignments apart. This dual objective directly penalizes instruction-agnostic representations: no single query encoding can satisfy both constraints unless it genuinely conditions on the

semantic content of the instruction, since $q \oplus I_{orig}$ and $q \oplus I_{new}$ must retrieve *opposite* documents.

Data synthesis and quality control. We use *Qwen3-Next-80B-A3B-Instruct* (Qwen Team, 2025) and seed instances from the Promptriever data that contain at least one pre-existing instruction negative. For each seed, the model either generates a complementary instruction or abstains when a valid reversal is infeasible; only successful generations enter the synthetic pool. Synthetic samples replace an equal-sized portion of the comparison data rather than being added on top.

We define a generated sample as *usable* only when all four criteria hold: the new instruction (1) makes the designated instruction negative relevant, (2) excludes the original positive while retaining its topical relevance, (3) keeps all remaining instruction negatives excluded, and (4) is fluent and contains no meta-task language. The original audit judged all 100 inspected samples usable; a second annotator independently judged 98 of the same 100 usable. An additional cross-family automated audit with *Gemma-4-31B* (Gemma Team et al., 2026) judged the intended polarity reversal correct for 979 of 1,000 inspected samples. These are post-hoc quality audits rather than training-time labels or a generator ablation. The complete synthesis prompt and validation rubric appear in Appendix A and Appendix B.

3 Experimental Setup

Backbone Models We adopt *gte-multilingual-mlm-base* (Zhang et al., 2024) (305M parameters) as our primary encoder, initialized from our own contrastively pretrained checkpoint trained on 1.41 billion unsupervised query-document pairs from C4 (Raffel et al., 2020), mC4 (Xue et al., 2021), CC News, and multilingual Wikipedia. To assess cross-backbone generalizability, we additionally train *bge-m3-retromae* (Chen et al., 2024) under the same data configurations and report results in Table 2.

Training data and controls. We separate a 480k-example diagnostic regime from the proposed 880k-example full-budget regime:

- *Ins-orig*: $\sim 480k$ original instruction-conditioned samples.
- *Ins-para* (control): $\sim 240k$ original samples and $\sim 240k$ label-preserving paraphrases generated from the same seed instances and with

the same generator as Ins-syn.

- *Ins-syn* (diagnostic): $\sim 240k$ original samples and $\sim 240k$ polarity-reversed samples, size-matched to Ins-orig and Ins-para. This isolates the effect of reversal when synthetic samples substitute for original contexts.
- *All-orig* (joint control): $\sim 440k$ instruction-conditioned and $\sim 440k$ non-instruction samples, implementing the data composition of Promptriever-style joint training (Weller et al., 2025b) with our backbone and optimization protocol.
- *All-syn* (proposed): the same $\sim 440k$ original instruction-conditioned samples as All-orig plus $\sim 440k$ polarity-reversed samples in place of its non-instruction half. This is our recommended full-budget recipe.

Training Details We use a contrastive training framework with 30 hard negatives per query, including 1–3 instruction negatives. We optimize with InfoNCE loss (van den Oord et al., 2019) with temperature $\tau = 0.02$. For both encoders, the query and instruction are concatenated before encoding, while documents are encoded independently. The maximum sequence length is 512 tokens for both queries and documents. All configurations use the same training protocol to ensure fair comparisons.

Evaluation benchmarks. We distinguish three evaluation dimensions. For *instruction sensitivity*, FollowIR (Weller et al., 2025a) and the InfoSearch (Zhou et al., 2025) length and keyword subsets report p -MRR, which measures whether ranks move in the correct direction when relevance-changing instructions are paired. For *instruction-conditioned ranking effectiveness*, we report the official FollowIR Score, InfoSearch nDCG@5, and MAIR (Sun et al., 2024) nDCG@10 on IFEval (Zhou et al., 2023) and InstructIR (Oh et al., 2024). For *standard retrieval*, we evaluate plain queries without instructions on FEVER, MS MARCO, Touche2020, and TREC-COVID from BEIR (Thakur et al., 2021), reporting nDCG@10. Finally, we form a diagnostic FollowIR variant by prompting *Qwen3.6-35B-A3B* (Qwen Team, 2026) to paraphrase each instruction without changing its intended semantics, and compare aggregate Score with the official instructions. Full metric and diagnostic details are in the appendix.

4 Results and Analysis

Proposed full-budget recipe. The primary comparison is All-syn against the size-matched All-orig joint control. On gte, All-syn raises FollowIR p -MRR from 5.27 to 8.30, an absolute gain of 3.03 points (+57.5%; 95% CI [2.11, 3.57], $p < .001$), while Score increases from 20.85 to 21.38. It also changes InfoSearch behavior from strongly negative to positive p -MRR and improves both MAIR subsets. The result is not specific to gte: on bge-m3, FollowIR p -MRR increases from 8.84 to 13.92, again +57.5% (difference 5.08; 95% CI [4.77, 5.44], $p < .001$), while Score increases from 20.69 to 20.99.

Ranking effectiveness and standard retrieval.

The effectiveness metrics qualify the sensitivity results rather than treating p -MRR as the sole objective. As Table 3 shows, All-syn improves InfoSearch nDCG@5 over All-orig on both length (73.65 to 75.98) and keyword (58.40 to 61.39), yielding the best average among the four main configurations. With plain BEIR queries and no instructions, All-syn is higher on three of four evaluated datasets and changes the average from 59.42 to 59.84. Given the small average difference, we interpret this as no observed average degradation on these datasets, not as evidence of a universal standard-retrieval improvement.

What does polarity reversal add? The size-matched Ins-para control isolates label reversal from additional LLM-generated instruction text. Its FollowIR p -MRR (5.04) remains close to Ins-orig (5.21), as do InfoSearch length/keyword p -MRR (3.05/3.28 vs. 4.06/2.06) and MAIR IFEval (31.11 vs. 32.14). Under the same generator, seed instances, and budget, Ins-syn instead reaches 7.57, 9.02/5.61, and 36.13. The comparison supports polarity reversal as the key factor behind the sensitivity gains relative to this paraphrase alternative. Both Ins-para and Ins-syn reduce FollowIR Score relative to Ins-orig, indicating that the diagnostic regime’s effectiveness loss is associated with substituting original contexts, rather than uniquely with reversal. The full-budget All-syn recipe avoids that substitution trade-off.

Consistency, boundaries, and robustness. Both matched comparisons improve FollowIR p -MRR on Robust04, News21, and Core17 (Table 5). Table 6 shows that all four aggregate gains are significant.

Training data	FollowIR		InfoSearch (p-MRR)		MAIR (nDCG@10)	
	p-MRR ↑	Score ↑	Length ↑	Keyword ↑	IFEval ↑	InstructIR ↑
<i>480k-example diagnostic configurations</i>						
Ins-orig	5.21	21.33	4.06	2.06	32.14	89.16
Ins-para (control)	5.04	20.58	3.05	3.28	31.11	88.72
Ins-syn (diagnostic)	7.57 [†]	19.73	9.02	5.61	36.13	87.97
<i>880k-example full-budget configurations</i>						
All-orig (joint control)	5.27	20.85	-23.22	-49.65	24.33	85.54
All-syn (proposed)	8.30[†]	21.38	31.91	12.13	34.08	90.74

Table 1: Main results on *gte-multilingual-mlm-base*. All-orig follows a Promptriever-style 50/50 joint-data recipe; All-syn is our proposed size-matched replacement. Ins-syn and Ins-para are diagnostic controls. Score is the macro-average of the official FollowIR effectiveness metrics. Negative *p*-MRR indicates movement opposite to instruction-defined relevance. [†]: $p < .001$ versus the corresponding -orig configuration under a query-level paired bootstrap test.

Training data	FollowIR		InfoSearch (p-MRR)		MAIR (nDCG@10)	
	p-MRR ↑	Score ↑	Length ↑	Keyword ↑	IFEval ↑	InstructIR ↑
<i>480k-example diagnostic configurations</i>						
Ins-orig	9.40	22.26	19.00	4.18	27.12	90.40
Ins-syn (diagnostic)	11.47 [†]	19.76	28.64	48.42	29.37	89.53
<i>880k-example full-budget configurations</i>						
All-orig (joint control)	8.84	20.69	-27.08	-62.04	25.68	90.67
All-syn (proposed)	13.92[†]	20.99	40.15	49.62	27.07	90.80

Table 2: Results on *bge-m3-retromae*. Roles and notation follow Table 1. The full-budget *p*-MRR improvement remains significant on this backbone.

The per-subset effectiveness results are less uniform: for example, All-syn slightly lowers Core17 Score while improving the macro-average, and Ins-syn lowers Score on all three subsets. We therefore distinguish instruction sensitivity from ranking effectiveness throughout. Finally, on paraphrases prompted to preserve the meaning of FollowIR instructions, the aggregate Score changes by at most 0.16 across the three tested configurations (Table 7). This provides evidence of aggregate stability under the tested rewordings, although it does not by itself establish per-query rank invariance.

Cross-backbone generalization. Table 2 reproduces the two controlled patterns on a 560M-parameter encoder: Ins-syn improves sensitivity but loses Score in the substitution diagnostic, whereas the proposed All-syn configuration improves every reported aggregate metric over its size-matched joint control. Thus the polarity-reversal benefit does not disappear on the larger of the two tested backbones.

5 Discussion

Polarity reversal versus instruction-based negatives. Promptriever (Weller et al., 2025b) and

InF-IR (Zhuang et al., 2025) show the value of instruction-tied negatives, but use each negative under a fixed relevance definition. Polarity reversal instead makes that document conditionally relevant under a complementary facet. The new view remains tied to an explicit label inversion over the same query-document set, rather than adding detached linguistic variation. The Ins-para control supports this distinction empirically: label-preserving paraphrases do not reproduce the sensitivity gains of reversed labels.

Instruction signal and data composition. The matched Ins- and All- comparisons expose different substitution effects. Replacing half of the original instruction contexts with reversed views sharpens sensitivity but reduces effectiveness, whereas replacing the non-instruction half at the 880k budget improves sensitivity without reducing the evaluated aggregate effectiveness. A descriptive 100/0–75/25–50/50 interpolation shows monotonically lower InfoSearch *p*-MRR as the non-instruction fraction increases (Table 8). Because the reused endpoints differ in total data size, this analysis is not a controlled causal estimate of mixture ratio. The stronger conclusion therefore comes from the matched 880k comparison.

(a) InfoSearch nDCG@5				(b) BEIR nDCG@10, plain queries					
Training data	Length	Keyword	Avg.	Training data	FEVER	MSM.	Touche	COVID	Avg.
Ins-orig	74.97	61.26	68.12	All-orig (joint)	91.33	43.65	25.42	77.27	59.42
Ins-syn (diagnostic)	76.75	59.92	68.34	All-syn (proposed)	91.52	43.85	26.88	77.12	59.84
All-orig (joint control)	73.65	58.40	66.03						
All-syn (proposed)	75.98	61.39	68.69						

Table 3: Additional ranking-effectiveness results on the gte backbone. All-syn improves its size-matched joint control on both instruction-conditioned InfoSearch subsets. On four evaluated standard-retrieval datasets, it is higher on three, slightly lower on TREC-COVID, and comparable on average.

6 Conclusion

We introduced dual-view training, which synthesizes complementary instructions that reverse relevance labels over fixed document pairs. At a matched full-data budget, the proposed recipe improves instruction sensitivity across two backbones while preserving the evaluated aggregate instruction-conditioned and standard-retrieval effectiveness. A paraphrase control indicates that label inversion, rather than extra generated wording alone, supplies the key training signal. The method changes only the supervision data and requires no architectural modification.

Limitations

Seed data requirements. Our synthesis pipeline starts from training examples that already contain instruction negatives. This matches several recent instruction-following retrieval resources, but applying the same procedure to other corpora may require first mining or annotating comparable negatives.

Generator and validation scope. We use one synthesis model, one prompt, and one seed corpus; neither the paraphrase control nor the cross-family quality judge is a generator or prompt ablation. The human and automated audits estimate validity on inspected samples but do not guarantee that all synthetic examples are correct. Generator swaps, prompt perturbations, and a taxonomy of unsuccessful reversals remain important future work.

Statistical scope. Our paired bootstrap tests quantify uncertainty from evaluation queries, not variation from training randomness. We did not repeat fine-tuning across random seeds. The paraphrase robustness diagnostic compares aggregate scores and can conceal offsetting per-query changes; a larger human-verified paraphrase set with per-query rank-stability measures would provide stronger evidence.

Models, baselines, and evaluation scope. We test 305M- and 560M-parameter encoders, three instruction-following suites, and four BEIR datasets on one backbone. These results should not be extrapolated to larger encoders, other languages, or deployment distributions without further study. All-orig faithfully instantiates Promptriever-style data composition within our protocol, but is not the original Promptriever model. We also do not retrain InF-IR head-to-head because its data scale and negative construction differ substantially; broader matched comparisons remain future work.

References

- 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.
- Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026. [Gemma 4 technical report](#). *Preprint*, arXiv:2607.02770.
- Hanseok Oh, Hyunji Lee, Seonghyeon Ye, Haebin Shin, Hansol Jang, Changwook Jun, and Minjoon Seo. 2024. [INSTRUCTIR: A benchmark for instruction following of information retrieval models](#). *Preprint*, arXiv:2402.14334.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Qwen Team. 2026. [Qwen3.6-35B-A3B: Agentic coding power, now open to all](#).
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text](#)

- [transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. [One embedder, any task: Instruction-finetuned text embeddings](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.
- Weiwei Sun, Zhengliang Shi, Wu Jiu Long, Lingyong Yan, Xinyu Ma, Yiding Liu, Min Cao, Dawei Yin, and Zhaochun Ren. 2024. [MAIR: A massive benchmark for evaluating instructed retrieval](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14044–14067, Miami, Florida, USA. 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*, volume 1.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. [Representation learning with contrastive predictive coding](#). *Preprint*, arXiv:1807.03748.
- Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2025a. [FollowIR: Evaluating and teaching information retrieval models to follow instructions](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11926–11942, Albuquerque, New Mexico. Association for Computational Linguistics.
- Orion Weller, Ben Van Durme, Dawn Lawrie, Ashwin Paranjape, Yuhao Zhang, and Jack Hessel. 2025b. [Promptriever: Instruction-trained retrievers can be prompted like language models](#). In *International Conference on Learning Representations*, volume 2025, pages 17660–17683.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. [mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models](#). *Preprint*, arXiv:2311.07911.
- Jianqun Zhou, Yuanlei Zheng, Wei Chen, Qianqian Zheng, Zeyuan Shang, Wei Zhang, Rui Meng, and Xiaoyu Shen. 2025. [Beyond content relevance: Evaluating instruction following in retrieval models](#). In *International Conference on Learning Representations*, volume 2025, pages 84965–84996.
- Yuchen Zhuang, Aaron Trinh, Rushi Qiang, Haotian Sun, Chao Zhang, Hanjun Dai, and Bo Dai. 2025. [Towards better instruction following retrieval models](#). *Preprint*, arXiv:2505.21439.

A The Prompt Template for Data Synthesis

The following box shows the full prompt template used for polarity-reversed instruction synthesis. Template variables (in monospace) are populated per instance.

Goal

Create a new synthetic instruction that **reverses** the original relevance judgment **only** for the specified passages:

- The **original positive passage** (P^+) must become an **instruction negative** under the new instruction (i.e., relevant to the pure query but **irrelevant** once the instruction is applied).
- The **specific instruction negative** (N^*) must become the **new positive** (i.e., relevant to the pure query **and** to the query+instruction).
- **All remaining instruction negatives** ($N_1 \dots N_k$) must **remain instruction negatives**.

You must not change the query or any passage content. Only write a new instruction.

Inputs

- query (string)
- original_instruction (string)
- positive_passage = P^+ (string)
- specific_instruction_negative = N^* (string)
- remaining_instruction_negatives = $N_1 \dots N_k$ (array; may be empty)

Output

You should reason step by step, and the final answer should be in the following XML format:

```
<answer>
<new_instruction>[your new instruction]</new_instruction>
</answer>
```

If you think this task is too hard to achieve, you should simply return `<answer>None</answer>`.

Definitions

- **Relevant to the pure query**: reasonably satisfies the user’s intent without any extra instruction.
- **Instruction negative**: relevant to the pure query, but **excluded by the instruction** (e.g., by scope, geography, timeframe, format, source constraints, audience level).

Method

- 1) *Profile passages*
 - Identify attributes of P^+ (domain, geography, timeframe, audience, medium/format, methodology, sources, constraints).
 - Identify attributes of N^* that **distinguish it from** P^+ .
 - Skim each N_i to note attributes you must **keep excluded**.
- 2) *Choose reversal levers*

Craft a new instruction that:

 - **Positively selects** N^* ’s attributes (so N^* becomes positive), and

- **Excludes P^+** via ≥ 1 **hard, objective constraint** (so P^+ becomes an instruction negative), while
- **Does not inadvertently admit** any N_i (keep them instruction negatives).

Useful levers: domain narrowing, region, timeframe/recency, audience level, style/format (e.g., “equations only”, “bulleted checklist”), methodology/evidence type, required artifacts (e.g., runnable code in a specific language), explicit exclusions (e.g., “exclude major capitals”, “exclude beach/island topics”).

3) Diversity requirement

Ensure the new instruction’s **format and perspective** differ from the original_instruction (e.g., switch voice, deliverable type, constraint style). Avoid trivial rewording.

4) Sanity checks

- Would P^+ be filtered out by the new constraints? If not, tighten them.
- Is N^* clearly included by the positive selectors? If not, pivot constraints toward N^* ’s attributes.
- Do all N_i still get excluded? If any slip in, add explicit exclusions or tighten scope.

5) Conciseness

The **new instruction must be one or two sentences**, imperative, concrete, and unambiguous.

Guardrails

- Do **not** reference passage IDs or this meta-task; exclude or include by **attribute** only.
- Do **not** modify the query or passages.

Checklist (before you output)

- P^+ becomes an instruction negative.
- N^* becomes positive.
- All N_i remain instruction negatives.
- Instruction is concise (≤ 2 sentences) and **diverse** vs. the original_instruction.
- No meta-task language or passage IDs appear in the instruction.

Now perform the task with the following data:

- Query: {{ query }}
- Original instruction: {{ original_instruction }}
- Original positive passage: {{ original_positive }}
- Specific instruction negative passage: {{ specific_instruction_negative }}
- All remaining instruction negative passages:


```
{% for negative in remaining_negatives %}
Negative passage {{ loop.index }}: {{ negative }}
{% endfor %}
```

B Synthetic-Data Quality Verification

Operational definition. We treated a generated instruction as usable only if all four conditions held jointly:

1. the designated instruction negative became relevant under the new instruction;
2. the original positive remained topically related but became excluded by the new instruction;
3. all remaining instruction negatives stayed excluded; and
4. the instruction was fluent, concrete, and free of passage identifiers or meta-task language.

A sample failed the audit if any criterion was not satisfied.

Quality audit	n	Pass rate
Initial human inspection	100	100.0%
Second independent human inspection	100	98.0%
Cross-family automatic polarity audit	1,000	97.9%

Table 4: Post-hoc checks of synthetic-data quality. The automatic check uses Gemma-4-31B and targets polarity consistency rather than the full human rubric.

Human and automatic audits. The second annotator independently applied the same four-criterion rubric to the same 100 examples and judged 98 usable. We report this usable rate rather than a chance-corrected agreement statistic because the initial inspection assigned the same label to all 100 examples. Separately, *Gemma-4-31B* (Gemma Team et al., 2026), from a different model family than the generator, received the query, both instructions, the positive, the designated instruction negative, and the remaining negatives. It confirmed the intended polarity assignment in 979 of 1,000 inspected examples. The automatic audit is supporting evidence about the evaluated outputs; it is not a substitute for human assessment and does not test robustness to the choice of synthesis generator.

Use in training. Both audits were conducted after data construction and did not supply training labels or filter successful generations. The synthesis prompt permits an explicit None response when reversal is infeasible; such abstentions are not treated as generated training examples.

C Experimental Details

Training. The contrastive pre-training checkpoint used to initialize our primary encoder was trained on 8 clusters, each containing 8 NVIDIA H200 GPUs. For all fine-tuning experiments, we used one cluster of 8 H200 GPUs. We trained for 1 epoch with a batch size of 32 per GPU, resulting in a total batch size of 256. The learning rate was set to 3×10^{-5} with a linear decay schedule and a warmup period of 10% of the total training steps. Each training run took approximately 12 hours to complete.

Paraphrase training control. Ins-para uses the same Qwen generator, seed instances, original/synthetic split, total sample count, and fine-tuning protocol as Ins-syn. The generated instruction paraphrases the original while keeping all relevance labels unchanged; Ins-syn instead changes

the instruction and reverses the designated document labels. Thus the comparison holds generated-text exposure fixed while changing the supervision relation.

Standard retrieval. For BEIR, we evaluate All-orig and All-syn on FEVER, MS MARCO, Touche2020, and TREC-COVID with plain queries and no appended instruction. We report nDCG@10 and an unweighted mean over the four datasets. All retrieval and evaluation settings are shared between the two configurations.

Prompted semantic-preserving paraphrases. We prompt *Qwen3.6-35B-A3B* (Qwen Team, 2026) to rewrite each FollowIR instruction without changing its intended semantics. Queries, documents, and official relevance labels remain fixed. We report the official Score on this separate diagnostic variant; all primary FollowIR results use the original instructions.

Statistical testing. For each of the four principal FollowIR comparisons, we align outputs by query and perform a paired bootstrap with 10,000 query-level resamples. Each replicate samples queries with replacement while retaining the two systems’ paired outputs and recomputes the aggregate p -MRR difference. The reported intervals and p -values are derived from the empirical bootstrap distribution. This procedure measures uncertainty from the evaluated query sample, not variation caused by retraining.

Mixture sensitivity. We additionally evaluate a 75/25 instruction/non-instruction configuration between the existing 100/0 Ins-orig and 50/50 All-orig endpoints. Since the reused endpoints have different total training-set sizes, this is a descriptive sensitivity analysis rather than a fixed-budget causal estimate of mixture ratio.

D Additional Results

Behavioral boundary. The subset results rule out a gain driven by only one collection, but also expose limitations. In the diagnostic regime, Ins-syn lowers Score on all three subsets. In the proposed full-budget regime, All-syn improves Robust04 and News21 Score but lowers Core17 from 19.69 to 19.14. A qualitative inspection of label-flipped documents found both expected directions of movement: newly relevant documents moved upward, while newly excluded documents moved

downward. We treat this as behavioral context rather than proof that the model is free of surface heuristics, because a complete quantitative failure taxonomy is not available.

E Evaluation Metrics

Pairwise MRR (p -MRR). We use p -MRR to measure whether a model changes its ranking in the correct direction when the instruction changes. For a document whose relevance flips across two paired instructions, let r^+ be its rank under the instruction where it is relevant and r^- be its rank under the instruction where it is not relevant. We define reciprocal ranks $RR^+ = 1/r^+$ and $RR^- = 1/r^-$, with $RR = 0$ if the document is not retrieved. The per-document score is

$$p\text{-MRR}(d) = \begin{cases} 1 - \frac{RR^-}{RR^+}, & \text{if } RR^+ \geq RR^- \\ \frac{RR^+}{RR^-} - 1, & \text{otherwise.} \end{cases}$$

Positive values indicate that the model ranks the document higher when it is relevant than when it is instruction-excluded, negative values indicate the opposite behavior, and zero indicates no rank change. Following the benchmark protocol, we average over flipped documents within each query and then macro-average over queries. Reported p -MRR values are multiplied by 100.

FollowIR Score. FollowIR reports both p -MRR and an effectiveness Score. The Score is the arithmetic mean of the official metrics across the three subsets: MAP@1000 for Robust04 and Core17 and nDCG@5 for News21. It measures instruction-conditioned ranking effectiveness, while p -MRR isolates sensitivity to instruction-induced relevance changes.

MAP@1000. For a query q , average precision at cutoff 1000 is

$$AP@1000(q) = \frac{\sum_{k=1}^{1000} P@k(q) rel_k}{\min(R_q, 1000)},$$

where R_q is the number of relevant documents for q , $P@k(q)$ is precision among the top- k retrieved documents, and rel_k indicates whether the document at rank k is relevant. MAP@1000 is the mean of $AP@1000(q)$ over queries.

nDCG@K. We report nDCG@5 for InfoSearch and the News21 component of FollowIR, and nDCG@10 for MAIR and BEIR. For a ranked list

Training data	Robust04		News21		Core17	
	<i>p</i> -MRR	Score	<i>p</i> -MRR	Score	<i>p</i> -MRR	Score
Ins-orig	2.19	19.56	3.90	26.52	9.55	17.91
Ins-syn (diagnostic)	5.34	18.34	5.31	24.56	12.05	16.28
All-orig (joint control)	3.58	16.61	5.54	26.26	6.68	19.69
All-syn (proposed)	4.93	16.96	7.25	28.04	12.72	19.14

Table 5: Per-subset FollowIR results on gte. Score is MAP@1000 for Robust04 and Core17 and nDCG@5 for News21. Sensitivity gains occur on every subset, while the effectiveness changes are not uniformly positive.

Comparison	Δ <i>p</i> -MRR [95% CI]	<i>p</i> -value
Ins-syn vs. Ins-orig (gte)	+2.36 [2.04, 2.51]	< 0.001
All-syn vs. All-orig (gte)	+3.03 [2.11, 3.57]	< 0.001
Ins-syn vs. Ins-orig (bge-m3)	+2.07 [1.48, 2.88]	< 0.001
All-syn vs. All-orig (bge-m3)	+5.08 [4.77, 5.44]	< 0.001

Table 6: Query-level paired bootstrap results on FollowIR (10,000 resamples).

Training data	Original	Paraphrased	Abs. Δ
Ins-orig	21.33	21.35	0.02
All-orig	20.85	20.94	0.09
All-syn	21.38	21.22	0.16

Table 7: Aggregate FollowIR Score under official instructions and paraphrases prompted to preserve their meaning. Abs. Δ is the absolute difference between aggregate scores, not the mean per-query absolute change.

with graded relevance g_k , we compute

$$DCG@K = \sum_{k=1}^K \frac{2^{g_k} - 1}{\log_2(k + 1)},$$

$$nDCG@K = \frac{DCG@K}{IDCG@K},$$

where $IDCG@K$ is the maximum possible $DCG@K$ for the query. Values are averaged over queries within each benchmark subset.

Benchmark-specific reporting. For each InfoSearch subset, we report both *p*-MRR and nDCG@5; their nDCG average is the unweighted mean of length and keyword. For MAIR, we report nDCG@10 separately on IFEval and InstructIR. BEIR uses plain queries and nDCG@10. All reported metrics are higher-is-better.

Inst.:non-inst.	F- <i>p</i> MRR	F-Score	Info-L	Info-K
100 : 0 (Ins-orig)	5.21	21.33	4.06	2.06
75 : 25	5.33	21.01	-1.04	0.08
50 : 50 (All-orig)	5.27	20.85	-23.22	-49.65

Table 8: Descriptive sensitivity to instruction/non-instruction composition on gte. F and Info denote FollowIR and InfoSearch; L/K are the length/keyword subsets. InfoSearch *p*-MRR decreases monotonically across the evaluated configurations, whereas FollowIR *p*-MRR does not. Because the reused endpoints differ in total size, this is not a fixed-budget ratio ablation.

F License Statement

Training and synthetic data. Our seed data is derived from the Promptriever instruction-training data (Weller et al., 2025b), an augmented version of MS MARCO. The Promptriever data card states that the dataset follows MS MARCO’s non-commercial research terms and the applicable Llama 3 and GPT-4o terms for its generated annotations. We use the data only for research, provide attribution to the original creators, and do not redistribute the original Promptriever/MS MARCO passages. The polarity-reversed and paraphrased instructions generated in this work are derived from Promptriever examples and remain subject to the applicable downstream-use constraints of the source data.

Evaluation benchmarks. We evaluate on FollowIR (Weller et al., 2025a), InfoSearch (Zhou et al., 2025), MAIR (Sun et al., 2024), and four datasets distributed through BEIR (Thakur et al., 2021). We use all benchmarks only for evaluation, follow the licenses or terms specified by their maintainers and underlying datasets, and do not redistribute benchmark contents. The paraphrased FollowIR diagnostic changes only instruction wording in a derived evaluation variant; the official benchmark and relevance labels remain unchanged.

Models and software. We use *Qwen3-Next-80B-A3B-Instruct* (Qwen Team, 2025) for training-

instruction synthesis, *Gemma-4-31B* (Gemma Team et al., 2026) for the automatic quality audit, and *Qwen3.6-35B-A3B* (Qwen Team, 2026) for the robustness diagnostic. These model families are released under Apache License 2.0. The encoder backbones *gte-multilingual-mlm-base* (Zhang et al., 2024) and *bge-m3-retromae* (Chen et al., 2024) are released under Apache 2.0 and MIT, respectively. All models and software are used in accordance with their applicable terms.

G GenAI Usage Statement

AI use in research artifacts. We used *Qwen3-Next-80B-A3B-Instruct* (Qwen Team, 2025) to synthesize the polarity-reversed training instructions and the label-preserving paraphrase control. For polarity reversal, the model received the query, original instruction, positive passage, one designated instruction negative, and remaining instruction negatives. The full reversal prompt is provided in [Appendix A](#).

AI-assisted quality audit. We used *Gemma-4-31B* (Gemma Team et al., 2026) as an automatic judge for a 1,000-example audit of the generated training instructions. Its outputs were used only as a diagnostic quality check and not as official benchmark relevance labels or additional training supervision. Human and automatic audit results are described in [Appendix B](#).

AI-assisted robustness evaluation. We used *Qwen3.6-35B-A3B* (Qwen Team, 2026) to create instruction paraphrases prompted to preserve meaning for a separate FollowIR diagnostic. The official queries, corpus, and relevance labels were not changed, and every primary FollowIR result uses the official instructions.

AI use in experiments. The retrieval backbones, training configurations, evaluation protocols, and reported numerical results were selected, run, and verified by the human authors. Generative AI did not create or modify official relevance judgments for FollowIR, InfoSearch, MAIR, or BEIR, nor did it produce the reported metric values.

AI use in writing. We used AI assistants, including OpenAI Codex, for editorial support, including grammar correction, sentence-level rephrasing, LaTeX cleanup, and wording suggestions. The human authors reviewed and verified the scientific claims, experimental design, analyses, citations, and con-

clusions. AI assistants were not used to invent citations or fabricate experimental results.

Responsibility. The authors take full responsibility for the paper, research artifacts, experimental results, and all text that may have received AI-assisted editorial support.