Dual-View Training for Instruction-Following Information Retrieval

Anonymous ACL submission

Abstract

Instruction-following information retrieval (IF-IR) studies retrieval systems that must not only find documents relevant to a query, but also obey explicit user constraints such as required attributes, exclusions, or output preferences. However, most retrievers are trained primarily for semantic relevance and often fail to distinguish documents that match the topic from those that satisfy the instruction. We propose a dual-view data synthesis strategy based on polarity reversal: given a query, a document that is relevant under the instruction, and a hard negative that matches the query but violates the instruction, we prompt an LLM to generate a complementary instruction under which the two documents swap relevance labels. By presenting the same document pair under complementary instructions that invert their relevance labels, the training signal forces the retriever to reconsider the same candidate set through the instruction, rather than relying on fixed topical cues. On a 305M-parameter encoder, our method improves FollowIR p-MRR by 45%, surpassing general-purpose embedding models of comparable or larger scale. Through head-to-head comparisons at matched data budgets, we further show that data diversity and instruction supervision play complementary roles: the former preserves general retrieval quality, while the latter improves instruction sensitivity. These results highlight the value of targeted data synthesis for building retrieval systems that are both broadly capable and instruction-aware.

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

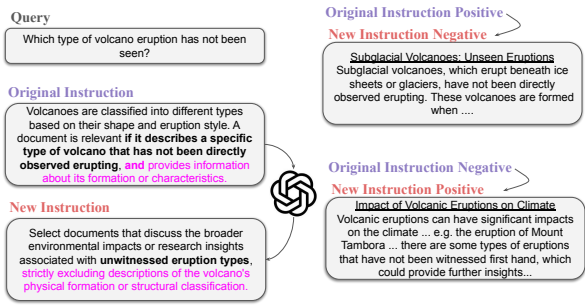

Figure 1: We synthesize new instructions that reverse the relevance polarity of existing documents, creating challenging samples that sharpen the retriever's sensitivity to instructional nuances.

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 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 synthesis strategy that improves FollowIR p-MRR by 45% on a 305M-parameter encoder, surpassing general-purpose embedding models of comparable or larger scale. (2) Through head-to-head comparisons at matched data budgets, we identify a fundamental tension in IF training: data diversity sustains general retrieval quality, while instruction supervision drives IF capability, and mixing non-instruct data without sufficient instruction signal can severely compromise the latter. (3) We show that dual-view synthesis resolves this tension, simultaneously improving instruction sensitivity and general retrieval at equal data budget. All findings are validated across two encoder backbones.

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 Setup We employ *Qwen3-Next-80B-A3B-Instruct* (Qwen Team, 2025) as the backbone LLM for our data synthesis pipeline. We construct our seed dataset by selecting instances from the promptriever data that contain at least one pre-existing instruction negative. For each of these data points, we generate a new instruction that reverses the roles of the positive document and the instruction negative. This process yields one complementary training sample per original instance. In our controlled experiments, synthetic samples substitute for an equal-sized portion of the training set rather than being added on top, enabling size-matched comparisons across all configurations. After synthesis, one annotator manually checked 100 data points and found all sampled synthetic instructions usable. Thus, no additional filtering was conducted. The specific prompt template used for this generation is provided in Appendix A.

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 We define four training configurations to isolate the effect of our synthesis method:

- *Ins-orig*: ∼480k samples drawn entirely from the original promptriever dataset (instruct data only).
- *Ins-syn* (ours): ∼240k original instruct samples + ∼240k synthetic samples from our polarity-reversal method (size-matched to Ins-orig).
- *All-orig*: ∼440k instruct samples + ∼440k non-instruct samples from the promptriever dataset.

- *All-syn* (ours): Replaces the non-instruct portion of All-orig with ∼440k synthetic polarity-reversed samples, paired with the same ∼440k original instruct samples (size-matched to All-orig).

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 evaluate on three benchmark suites: (1) **FollowIR** (Weller et al., 2025a), reporting p-MRR for instruction sensitivity and an aggregated Score for overall retrieval quality. The p-MRR metric quantifies instruction sensitivity by comparing a model's ranking under two paired instructions: one where a document is annotated as relevant and one where it is not. Positive p-MRR indicates the model correctly adjusts its ranking in response to instruction changes, while negative values indicate it ranks documents in the *opposite* direction. (2) **InfoSearch** (Zhou et al., 2025) *length* and *keyword* subsets, also reporting p-MRR; and (3) **MAIR** (Sun et al., 2024) IFEval (Zhou et al., 2023) and InstructIR (Oh et al., 2024) subsets, reporting nDCG@10.

4 Results and Analysis

Table 1 summarizes results across all benchmarks. We organize our analysis around two experimental comparisons that together reveal the interplay between instruction sensitivity and data diversity.

Instruct-only comparison: IF gains at the cost of general retrieval. In the size-matched comparison between Ins-syn and Ins-orig (∼480k samples each), our synthetic data yields large gains on the paired instruction-sensitivity metrics: FollowIR p-MRR increases from 5.21 to 7.57 (+45%), surpassing general-purpose models such as EmbeddingGemma-300M (5.61 p-MRR) (Vera et al., 2025) and Qwen3-Embedding-0.6B (5.09 p-MRR) (Zhang et al., 2025); InfoSearch p-MRR improves by +122% (length) and +172% (keyword); and MAIR IFEval rises from 32.14 to 36.13. However, FollowIR Score drops from 21.33 to 19.73, and MAIR InstructIR changes modestly from 89.16 to 87.97. Since Ins-syn replaces half of the original instruct samples with synthetic counterparts, the model sees fewer unique training contexts, which may account for this decline. This observation points to data diversity as a key factor in sustaining general retrieval performance, a hypothesis we test directly in the All- configurations below.

Mixed-data comparison: All-orig vs. All-syn. The All- configurations compare two strategies for scaling the training set to ∼880k while holding the instruct portion fixed (∼440k): supplementing with non-instruct data (All-orig) versus synthetic polarity-reversed data (All-syn). Across all benchmarks, All-syn outperforms All-orig. FollowIR p-MRR improves from 5.27 to 8.30, the highest among all configurations, and Score rises from 20.85 to 21.38. On InfoSearch, All-syn achieves positive p-MRR (31.91 length, 12.13 keyword) whereas All-orig falls into negative territory (-23.22, -49.65), indicating that the model with non-instruct supplementation contradicts instruction-defined relevance. MAIR metrics follow the same pattern: All-syn improves over All-orig on both subsets, reaching the best InstructIR score and a strong IFEval score.

The role of data diversity. These results, combined with the Ins- comparison, clarify the respective roles of data diversity and instruction supervision. The Ins- experiments show that dedicated instruction data drives IF capability, but replacing original samples with synthetic ones reduces diversity and costs general retrieval quality. The All- experiments reveal the converse: scaling with non-instruct data provides diversity but dilutes the instruction signal, severely degrading IF while offering only marginal general retrieval benefit over Ins-orig. All-syn achieves the best of both by providing instruction-conditioned training pairs at scale, simultaneously maintaining the data volume that sustains general quality and the instruction signal that drives IF capability. Notably, All-syn contains no non-instruct data, yet achieves the best Score across all configurations. This suggests that data volume, rather than source heterogeneity per se, is the primary driver of general retrieval quality, as long as individual training samples are sufficiently diverse in their query-document pairings.

Cross-backbone generalizability. Table 2 reports results on *bge-m3-retromae*, a stronger en-

Training data	FollowIR		InfoSearch (p-MRR)		MAIR (nDCG@10)	
	p-MRR ↑	Score ↑	Length ↑	Keyword ↑	IFEval ↑	InstructIR ↑
Ins-orig	5.21	21.33	4.06	2.06	32.14	89.16
Ins-syn (ours)	7.57	19.73	9.02	5.61	36.13	87.97
All-orig	5.27	20.85	-23.22	-49.65	24.33	85.54
All-syn (ours)	8.30	21.38	31.91	12.13	34.08	90.74

Table 1: Main results on instruction-following retrieval benchmarks. Ins-/All- denote instruct-only and mixed training regimes; -orig/-syn denote original and our augmented data. Score is the macro-average across three FollowIR subsets (MAP@1000 on two subsets; nDCG@5 on one). Negative p-MRR indicates the model *contradicts* instruction-defined relevance.

Training data	FollowIR		InfoSearch (p-MRR)		MAIR (nDCG@10)	
	p-MRR ↑	Score ↑	Length ↑	Keyword ↑	IFEval ↑	InstructIR ↑
Ins-orig	9.40	22.26	19.00	4.18	27.12	90.40
Ins-syn (ours)	11.47	19.76	28.64	48.42	29.37	89.53
All-orig	8.84	20.69	-27.08	-62.04	25.68	90.67
All-syn (ours)	13.92	20.99	40.15	49.62	27.07	90.80

Table 2: Results on *bge-m3-retromae*. Score follows Table 1; synthetic data improves p-MRR metrics and offsets non-instruct mixing degradation.

coder with a different pretraining strategy. Both experimental patterns replicate faithfully. In the Ins- comparison, Ins-syn improves the p-MRR metrics: FollowIR p-MRR rises from 9.40 to 11.47 (+22%), and InfoSearch length/keyword p-MRR rises from 19.00/4.18 to 28.64/48.42, while Score drops from 22.26 to 19.76, confirming the same IF/diversity trade-off. In the All- comparison, the same degradation pattern reappears with All-orig (-27.08 and -62.04 on InfoSearch), and All-syn again reverses it entirely (40.15 and 49.62), achieving the best FollowIR p-MRR of 13.92 while maintaining competitive Score (20.99). The magnitude of the keyword gains is considerably larger on bge-m3, suggesting that a stronger backbone amplifies the benefit of our polarity-reversal signal. The cross-model consistency confirms that both the synthesis method and the data mixing degradation are backbone-agnostic phenomena, reinforcing the generality of our findings.

5 Discussion

Polarity reversal versus instruction-based negatives. Promptriever (Weller et al., 2025b) and InF-IR (Zhuang et al., 2025) show that instruction-tied negatives outperform generic hard negatives, but still treat them as fixed failures under a single instruction. Polarity reversal instead turns each negative into a *conditionally relevant* document under a complementary facet, echoing LLM work that uses curated multi-faceted instructions to strengthen instruction following (Lou et al., 2024). Our setting adds a retrieval-specific constraint: the new facet is tied to an explicit label inversion over the same query-document set, keeping each added facet tied to the original ranking ambiguity rather than a detached paraphrase. Thus instruction diversity is not only a source of linguistic variation, but also a source of contrastive supervision across relevance definitions. The query encoder must resolve where I_{orig} and I_{new} diverge, rather than merely learning which documents one instruction excludes, so the signal targets instructional distinctions rather than instructional exclusions.

A gradient perspective on data mixing. The data mixing degradation explains why larger generic embedders can trail instruction-specialized models: non-instruct samples reward query-correlated retrieval and can overwhelm conditioning at a 50/50 mix. Instruction sensitivity therefore requires *consistent* supervision, consistent with InF-IR (Zhuang et al., 2025), which achieves competitive IF performance from $\sim$38k specialized triplets. Our synthesis supplies this signal at scale by converting instruction negatives into instruction-conditioned positives, preserving data volume while making every training pair reinforce instruction-aware representations. In short, dual-view training reconciles instruction signal for IF sensitivity with data diversity for general retrieval quality at the same budget.

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.

Evaluation scope. We focus on controlled comparisons across two encoder backbones and three instruction-following retrieval benchmark suites. Broader evaluations across additional retrieval domains and deployment settings would further clarify how the observed trends vary with task distribution.

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.

Renze Lou, Kai Zhang, Jian Xie, Yuxuan Sun, Janice Ahn, Hanzi Xu, Yu Su, and Wenpeng Yin. 2024. Muffin: Curating multi-faceted instructions for improving instruction-following. In *International Conference on Learning Representations (ICLR)*.

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.

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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *Preprint*, arXiv:1807.03748.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 70 others. 2025. EmbeddingGemma: Powerful and lightweight text representations. *Preprint*, arXiv:2509.20354.

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.

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. *Preprint*, arXiv:2506.05176.

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 ($N^\star$)** must become the **new positive** (i.e., relevant to the pure query **and** to the query+instruction).
- **All remaining instruction negatives ($N_1 \ldots N_k$)** 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` = $N^\star$ (string)
- `remaining_instruction_negatives` = $N_1 \ldots N_k$ (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 $N^\star$ 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** $N^\star$'s attributes (so $N^\star$ 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 $N^\star$ clearly included by the positive selectors? If not, pivot constraints toward $N^\star$'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.
- $N^\star$ becomes positive.
- All N_i remain instruction negatives.
- Instruction is concise ($\leq$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 Experimental Details

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.

C 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 a standard retrieval Score. The Score is the arithmetic mean of the official standard retrieval metrics across the three FollowIR subsets: MAP@1000 for the two MAP-based subsets and nDCG@5 for the nDCG-based subset. This metric reflects general retrieval quality under the provided instruction, 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. For MAIR we report nDCG@10, and for one FollowIR subset the official Score uses nDCG@5. For a ranked list 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. The final value is averaged over queries within each benchmark subset.

Benchmark-specific reporting. For InfoSearch length and keyword subsets, we report p-MRR using the same pairwise calculation above. For MAIR, we report nDCG@10 separately on the IFEval and InstructIR subsets. All metrics are higher-is-better.

D 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. MS MARCO is released under the Creative Commons Attribution 4.0 International License (CC BY 4.0), and the Promptriever data card additionally states that its LLM-generated instructions and generated instruction positives/negatives follow the applicable Llama 3 and GPT-4o model terms. We use the data in accordance with these terms, provide attribution to the original creators, and do not redistribute the original Promptriever/MS MARCO passages. The synthetic polarity-reversed instructions generated in this work are derived from Promptriever examples for research use and inherit the applicable downstream-use constraints of the source data.

Evaluation benchmarks. We evaluate on FollowIR (Weller et al., 2025a), InfoSearch (Zhou et al., 2025), and MAIR (Sun et al., 2024). InfoSearch is released under the Creative Commons Attribution 4.0 International License, and the MTEB-hosted FollowIR tasks provide license metadata where available; MAIR aggregates tasks from many source datasets. We use all benchmarks only for evaluation, follow the license or terms specified by their maintainers and underlying source datasets, and do not redistribute benchmark contents.

Models and software. We use *Qwen3-Next-80B-A3B-Instruct* (Qwen Team, 2025) for instruction synthesis under the 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. The training framework used for fine-tuning is released under Apache 2.0. All models and software are used in accordance with their respective licenses.

E GenAI Usage Statement

AI use in research artifacts. Generative AI is an explicit component of our proposed data synthesis method. Specifically, we used *Qwen3-Next-80B-A3B-Instruct* (Qwen Team, 2025) to synthesize polarity-reversed instructions from Promptriever examples. For each selected instance, the model was given the query, original instruction, positive passage, one selected instruction negative, and remaining instruction negatives, and was prompted to

produce a new instruction under which the selected instruction negative becomes relevant and the original positive becomes an instruction negative. The full prompt template is provided in Appendix A. We manually checked 100 generated examples and found all sampled examples usable; no additional filtering was conducted.

AI use in experiments. The generated instructions constitute the synthetic training data evaluated in this paper. Retrieval backbones, training configurations, evaluation benchmarks, metrics, and all reported numerical results were selected, run, and verified by the human authors. Generative AI was not used to modify benchmark test sets, assign evaluation labels, or produce reported metric values.

AI use in writing. We used AI assistants, including OpenAI Codex, for surface-level editorial support, including grammar correction, sentence-level rephrasing for clarity and concision, LaTeX cleanup, and wording suggestions. The human authors wrote and verified the scientific claims, experimental design, analyses, citations, and conclusions. AI assistants were not used to invent citations, fabricate results, or introduce unverified factual content.

Responsibility. The authors take full responsibility for the content of this paper, including all research artifacts, experimental results, and text that may have been drafted or edited with AI assistance.